DHQ: Digital Humanities Quarterly
2009
Volume 3 Number 2

Published Yet Never Done: The Tension Between Projection and Completion in Digital Humanities Research

Susan Brown <sbrown_at_uoguelph_dot_ca>, University of Guelph
Patricia Clements <patricia_dot_clements_at_ualberta_dot_ca>, University of Alberta
Isobel Grundy <isobel_dot_grundy_at_ualberta_dot_ca>, University of Alberta
Stan Ruecker <sruecker_at_ualberta_dot_ca>, University of Alberta
Jeffery Antoniuk <jeffery_dot_antoniuk_at_ualberta_dot_ca>, University of Alberta
Sharon Balazs <sbalazs_at_ualberta_dot_ca>, University of Alberta

Abstract

The case of the Orlando Project offers a useful interrogation of concepts like completion and finality, as they emerge in the arena of electronic publication. The idea of "doneness" circulates discursively within a complex and evolving scholarly ecology where new modes of digital publication are changing our conceptions of textuality, at the same time that models of publication, funding, and archiving are rapidly changing. Within this ecology, it is instrumental and indeed valuable to consider particular tasks and stages done, even as the capacities of digital media push against a sense of finality. However, careful interrogation of aims and ends is required to think through the relation of a digital project to completion, whether modular, provisional, or of the project as a whole.

When can a digital scholarly project be considered finally "done"? Perhaps never. Something done is past, irrevocable, requiring nothing more and indeed immune from further action. The case of the Orlando Project, a large-scale and longstanding digital humanities undertaking, reveals an arbitrariness, even a fictiveness or contradictoriness, to the notion of completion of the project as a whole or even of its major online product. Digital humanities projects are considerably more prone than traditional humanities undertakings to riding off into the sunset until the next installment rather than being laid to rest. "Doneness" circulates discursively within a complex and evolving scholarly ecology where new modes of digital publication are changing our conceptions of textuality, at the same time that models of publication, funding, and archiving are rapidly changing. Within this ecology, it is instrumental and indeed valuable (indeed, as Matt Kirschenbaum suggests here, highly satisfying) to consider particular tasks and stages done, even as the capacities of digital media push against a sense of finality. However, careful interrogation of aims and ends is required to think through the relation of a digital project to completion, whether modular, provisional, or of the project as a whole. 1

Projection and Experimentality

In the digital humanities we often organize undertakings in terms of "projects", research endeavours that are probably, ideally, a collaborative enterprise "carefully planned to achieve a particular aim" (Oxford English Dictionary 2007, "project"). The emphasis is on the future, on the projected outcome and potential of the undertaking: projects, as the cognate verb "to project" suggests, are future-oriented. Some — an example would be the nora Project [nora] — last about as long as the money from a particular grant, but others — the Perseus Project and the Women Writers Project are examples — continue over many years and multiple grant funding cycles. A successful project is thus not necessarily geared to realizing a "particular aim". Perseus as "an evolving digital library" (http://www.perseus.tufts.edu/) situates its work in the vast scale of biological time; the WWP's aims are equally open-ended. These trajectories of digital humanities undertakings don't pin themselves to a specific end, but carry "A planned or proposed undertaking; a scheme, a proposal; a purpose, an objective" (Oxford English Dictionary 2007, "project") into the foreseeable future, with gusto. Such an orientation is actually at odds with the definition of a project in relation to particular aims. The success of these projects is not pegged on completion, but measured in other ways. 2

So there's clearly a lot that scholars involved with such projects want to do without being done, particularly insofar as being undone is compatible with disseminating materials to others and engaging in scholarly dialogue about them. At root is not only, as the introduction here suggests, a culture of perpetual prototypes that mitigates desire for closure, or funding structures that poorly support the "finishing" process for non-commercialized projects. It is also the very multi-faceted nature of much digital humanities research, which so often straddles the divide between content development and technological experimentation. This interplay between traditional humanities content and innovative methodologies means there is always more to be done. 3

The Orlando Project, with its aim of "producing the first full scholarly history of women's writing in the British Isles" (Brown et al. 2006, home page), is a long-term digital humanities project that is both done and yet not done. Unlike many electronic projects, the project held off making its major resource available until it was in quite a polished and complete state. *Orlando: Women Writers in the British Isles from the Beginnings to the Present* was published online by Cambridge University Press in June 2006. It is not a collection of primary texts, but a massive born-digital resource in literary history amounting now to almost 7.7 million words in the form of 1,206 detailed and often quite lengthy entries on writers' lives and writing careers, more than 13,000 independent chronology items, and 22,000 bibliographical records. Yet the project is far from done: its content and technical work continue. This paper explores the tension between projection and completion over this project's history to date as a means of considering that tension in relation to digital humanities research generally. 4

The project's cofounders (Susan Brown, Patricia Clements, and Isobel Grundy) were new to digital humanities research, so our notions of scholarly process and completion related to conventional print publications. As Claire Warwick has noted, the idea of what is "complete" or "publication-ready" in academic culture has emerged from a complex set of human factors relating to such matters as the attribution of credit by institutions and funding structures, as well as the conception of what is required intellectually for a product to be done [Warwick 2004, 368]. Such factors undoubtedly entered into how the Orlando Project was conceived. In our original funding application, we projected a single moment of completion at which the planned electronic history would be ready alongside several related print volumes of scholarship. We were fortunate to receive a Major Collaborative Research Initiative (MCRI) grant from the Social Sciences and Humanities Research Council of Canada (SSHRC), and embarked on the project as planned. And then things changed. 5

Working at the interface between humanities research questions and evolving digital methods means that projections about the trajectories of digital humanities work are less likely to be accurate than those of traditional scholarship. This is not to say that any research project may not run into snags or unforeseen delays - instances, particularly in the history of earlier scholarship, include Samuel Johnson's having to restart his dictionary mid-stream because he realized that working with small slips of paper would be better than the old technology of full sheets - but these are less often related to the methodology per se of the scholarly undertaking. In the case of Orlando, the ambition and experimentality of what we had undertaken on the technical side had a radical impact on the progress of the literary work with which it was interdependent, both because key researcher time was involved in the development of the custom tagset we developed and successively refined as a key component of our methodological experimentation, and because we had to build in-house production and delivery systems from scratch in ways that we had not anticipated. The risk of these sorts of impacts is endemic to methodologically experimental research of any kind, and particularly relevant to digital humanities work. Such impacts don't mean that the project is not pursuing its aims effectively, but they can have a major impact on anticipated timelines and perceptions of productivity, especially if the project has been articulated in relation to a particular aim or deliverable. 6

Modularity and Incrementalism

Digital humanists therefore need to plan and sequence with care their deliverables, which are important not only because our work must take objective form to be shared with our colleagues, but also because those are the ways in which we are accountable to the funding bodies that make our often costly work possible. Given the risk-oriented nature of experimental research, it is strategic to promise outcomes that are both multiple and modular. The Orlando Project struggled for funding in later stages as a result, we believe, of a project design that focused on a single, end-loaded monumental deliverable. 7

The big "ta-da!" moment of publication is a very common strategy, one followed by the Blake Archive in 1996 with its release of *The Book of Thel*, copy F, and then again in 1997 with further fanfare when it released the first SGML version of that text with additional functionality for users, and in 2008 by The Nineteenth-Century Serials Edition, which launched all six serials with a splash. The "ta-da" provides both that crucial sense of satisfaction and progress for the participants and a landmark achievement that constitutes important evidence of completion of at least a phase of the project for funding agencies. It has drawbacks, however. Focus on an end deliverable can obscure interim accomplishments. The Orlando Project's research plan was designed to proceed through a number of stages. Indeed, the milestones and the mid-term review required by the MCRI Program are examples of the kind of official part-done marker which, although it may not arise organically from the research needs or achievements and is imposed from the outside by the bureaucratic rules of another entity, can nevertheless be used by researchers as a spur to setting and meeting meaningful interim goals. Many of those markers were, however, internal to the project, which meant that they didn't provide the same kind of objective sense of progress that comes with public release. Externally, and for our funding agency, what registered was what we had not done, rather than what we had accomplished. An immense online "product" such as that Orlando promised from the outset does well then to be balanced by some objective, interim goals. In addition, the launch that suits a book does run somewhat counter to the ongoing life of many digital projects if it leaves people thinking that the project itself is finished. 8

Other projects have made their way into the world rather differently, in ways suggestive of ongoing curation. The Poetess Archive, for instance, began as a series of rather modest web pages that have grown over time in both scope and sophistication, moving from HTML into XML with a sophisticated search interface. Editorial ventures perhaps lend themselves particularly well to an incremental approach. The Brown University Women Writers Project (http://www.wwp.brown.edu), for instance, first transcribed and encoded texts, made printouts available and partnered with a publishing house, and released Renaissance Women Online. By the time Women Writers Online was made available by subscription, it was clear that, although it was a major event, it was part of a continuing project. Just as software projects typically put out numbered releases, which provide the triumphant moment of celebration while suggesting that more is yet to come, designing projects to incorporate such incrementalism by way of staged releases that mark phases of accomplishment or a number of discrete and in some way publishable deliverables, seems a particularly useful way to structure digital projects. 9

So it seems crucial to design digital humanities projects with a number of discrete and in some way publishable deliverables. Ideally, these should be modular, that is, functionally independent of one another. This means each has the potential for separate funding, can proceed on its own, and can provide a satisfying moment of completion. Modularity, however, suits some kinds of projects better than others. Both content and software systems often rely on the interrelation of various parts that can make it a challenge for one part to develop independently of the others. And even where a high degree of modularity is possible, modules usually need to be integrated at some point, so careful coordination to ensure eventual compatibility is still necessary, as well as an eventual convergence of module completion. 10

Various factors can work against modular publication. *Orlando*'s content structure was modular in form, composed as it was of author entries and chronological materials. Each of these theoretically could have been published as soon as they were "done". Yet doneness there was relative: there occurred a regular effect whereby the production of a new entry spurred significant improvement in several supposedly complete ones. An iterative process developed, not unlike the successive stages required in traditional humanities research, where the gradual accretion of knowledge slowly modifies the researcher's view and understanding of material. There was a strong sense both that the content work had to progress to a certain point of intellectual maturity, and that there were intellectual demands for a certain degree of coverage. We wouldn't be "done", for instance, without having completed the materials on Virginia Woolf or George Eliot. Because much feminist work has resisted the establishment of a small canon of female writers at the expense of others, such major writers needed to be situated in relation to less prominent contemporaries, and because we rejected a separatist understanding of literary history, we needed to include some male and international writers. Despite the apparent modularity of our content, we held off publishing until we had 1,149 entries completed. Thus, where scholarly content is concerned, a certain critical mass may be held necessary to establish scholarly confidence in the quality of a resources. Whether that threshold constitutes a single digital object, such as an edited text, or thousands of objects will vary. But it can work against a modular approach. Further revision is of course possible: the *Orlando* entries on Eliot and Woolf continue to be extended or revised at almost every update. In this sense, the digital done with its easy accommodation of incrementation is infinitely preferable to the printed done. But any project wishing to publish in stages will have to decide its initial content threshold according to the particular research goals of the project, criteria in the field for scholarly reliability, and user expectations. 11

Technical considerations constitute a further challenge to modular publishing, since a prototype is one thing and a debugged, multiple-browser-supporting, polished publication vehicle is another. We know that users are very easily put off by frustration in the use of new resources or tools, so publishing components that are unstable or poorly integrated may have a seriously negative impact. In the case of *Orlando*, our customized tagset required us to build a fairly complex XML delivery system, a task we had not anticipated in the mid-90s when TEI-SGML was emerging as a standard and XML was just over the horizon. Only a quite finished interface, we felt, stood a chance of convincing our core users from the technologically-resistant field of literary studies of the strengths of the markup into which the project had invested so much intellectual labour. *Orlando* offers users a range of affordances beyond that of looking up specific writers' entries, as the menu bar on the home page as it was at initial release (see Figure 1) makes clear. These extend to searching in quite precise ways on the more than 2 million semantic tags embedded in its literary-historical prose. To make the system's unique strengths apparent, we again needed a critical mass of materials to populate search results and showcase innovative features — such as the links screens that provide semantically-categorized access to mentions of writers across the textbase (see Figure 2). 12

Figure 1. 13

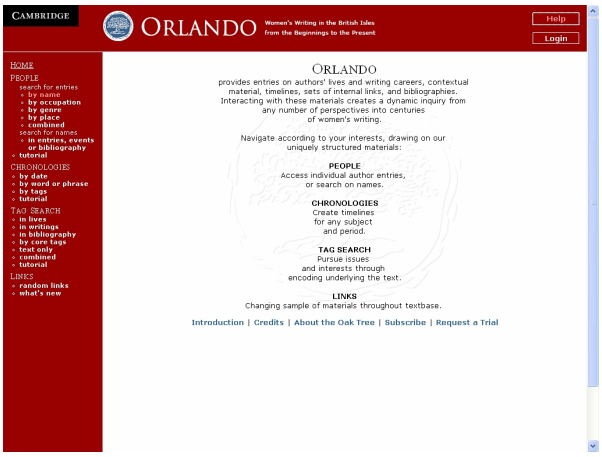


Figure 1: *Orlando* home page, showing the range of affordances in the left-hand menu.

Figure 2.

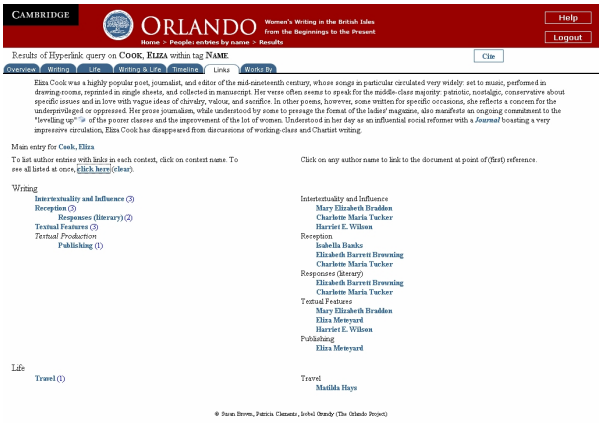


Figure 2: *Orlando* links screen for Eliza Cook. The right-hand column lists the available links to mentions of Cook in other entries, categorized by context according to the semantic context indicated by the XML tags.

Orlando shifted to a more staged publication model by uncoupling the electronic from the print publication, so that the former stands alone initially. Yet the textbase was published relatively complete. Thus, while structuring projects modularly is highly desirable for a range of reasons, truly modular publication may present challenges with respect to audiences from beyond the digital humanities community. Research domain, project conceptualization, and publication options are all crucial determinants of how "done" will be defined for a particular project. Project members need to arrive at a shared understanding of what constitutes an acceptable degree of intellectual maturity, critical mass of content, and technological finish at initial publication. This is particularly important since projects often seem to be judged by both funders and traditional humanities users according to their state at first release, as if they were a book. Once a first set of material is released, staged publication — such as the addition of new components, functionalities, or alternative interfaces — and incrementation — such as additions to or enhancement of existing content — become easier. However, in project planning, it seems strategically important for researchers to stress to funders the value of interim publications and subprojects, and generally not to allow a major deliverable to swallow up the identity of a project as a whole, so that the perception that the former is "done" does not carry with it a sense that the latter is also finished. Release or version numbers, or other ways of flagging the open-endedness of an electronic publication may be helpful in this regard. 15

Digital Textuality and Publication

Digital projects, if they aim to move beyond prototypes and court a mainstream humanities user community, need to recognize at the planning and budgeting stage the very high overhead involved in the development of delivery systems robust and usable enough to be considered in some sense finished.[1] We need to think through with our funding agencies not only how to sustain digital publications over the long haul, but also how to help projects with hugely valuable content leap that imposing hurdle from prototype to polished publication. At the same time, digitally published may not mean "done" in several respects. 16

Published is traditionally done, as David Sewell argues in his essay in this cluster. But published electronic projects don't get put on a shelf in a library. Being unconstrained by print materiality reinforces the arbitrariness of deciding that something is done in the sense of "complete", which is defined in the *Oxford English Dictionary* as "Having all its parts or members; comprising the full number or amount; embracing all the requisite items, details, topics, etc.; entire, full." Published may mean (provisionally) done without meaning complete, and there is of course a long tradition of encyclopedic print publications issuing a series of updates or supplements. Digital publication allows us to define done in terms of the kind of intrinsic completeness suggested by the *OED* rather than because we've reached an arbitrary limit (a deadline, a word length) related to print processes. In this sense, *Orlando*, though published, remains incomplete. Though all the items we considered requisite for initial publication are there, we remain aware of those figures, topics, approaches, and perspectives that demand inclusion in a "full" history of women's writing in the British Isles. Our contract with our publisher recognises the provisionality of our completion by stipulating for updates, as well as in the plan for the volumes of discursive history. We've increased and enhanced both content and functionality semi-annually since publication. 17

The "done" founded on digital publication is fragile in another sense because of the rapidity of technological change. The stability of book technology means that a book can be done and put to rest by both authors and publishers: even if it goes out of print, so long as copies endure in libraries they can continue to be used in perpetuity. But digital publications require more active support. Even if no technological enhancements are desired, for an electronic text to remain usable, it has to be stored somewhere in a form that is accessible to evolving technologies. This means it requires more active curation: even a quite straightforward web publication becomes unusable if it can't keep pace with browser releases. A new version of a project produced to migrate with current standards and practices is different from a second print edition in a number of respects. While both respond to a perception of continued demand for the product, the electronic migration is required to keep the resource accessible at all, and it does not supplement the first edition, which in the case of print will persist, but materially speaking supercedes or replaces it. This means that updates to electronic publications, while having a decided formal edge over errata slips or supplemental volumes, bear the additional burden of keeping the text in circulation. Being done with a digital publication may mean that the work disappears entirely from use. 18

The potential evanescence of a project's digital output creates pressures on the scholar, team, or publisher to keep it available. The academic community is still groping to discover how best to sustain digital publications over the long term. In the meantime, to meet even modest needs for technical migration and to keep content current, projects must continue to find funding, which can be challenging if a project is perceived as done as a result of publication. The Orlando Project, as part of its strategy of sustainability, licensed the textbase to the University of Alberta, and the University in turn sub-licensed it to Cambridge University Press. This arrangement created a revenue stream to help support the project's preparations for publication and its updates and ongoing activities. It also sustains Orlando's identity at its home institution and gives a broader constituency than the team members an interest in the project's success. This is important because, although like other ongoing projects Orlando has been able to obtain research funding for new initiatives, maintenance funding is a major challenge. 19

Part of the problem is in how we conceive of digital publications. Many ongoing digital publications should be understood by analogy with journals, for whom "done" can be applied to particular issues but not to the relevant research area. Continuing work despite previous publication is then part of the mandate, rather than the extraordinary burden it would seem in comparison with a book. The analogy applies only in part, because of course the entire text of a digital publication is fluid and subject to ongoing revision as that of a print journal is not. But it helps conceive more appropriately what "done" might mean for a lot of digital projects, with their capacity to increment and to migrate both technologically and with their field, just as does the analogy of the library for the Perseus Project. 20

Indeed, from a theoretical perspective, an electronic publication will arguably never be "done" precisely because of the nature of electronic textuality. Print texts are susceptible (as indeed were manuscripts and printed texts) to all sorts of repurposing, from reissue through quotation and anthologizing, to reprinting or incorporating in works of graphic art. In a digital environment, this aspect of textuality is greatly intensified by the ease with which one can "sample" texts, and the ability to separate content from presentation in digital formats means that entire works can be readily reformed or deformed. To take a familiar digital activity as an example, textual editing in an electronic environment must be reconceived as involving several different modes of editing. A TEI-conformant XML edition can form the basis of other quite divergent editions, such as an intentionalist rather than a genetic or "fluid-text" edition such as the Rotunda Press edition of Melville's *Typee* [Bryant 2006]. Scholars will increasingly be able to build on existing electronic texts, restructuring or adding to them, or recombining them with new content to produce new texts. In a radical extension of earlier forms of textuality, the possibility that an electronic text will continue to morph, be reproduced, and live on in ways quite unforeseen by its producers makes "done" to an extent always provisional. 21

Archives

The fact that electronic texts are not static leads to the thorny issue of archiving them, surely a marker of some kind of doneness. For although the "digital archive" is used loosely to refer to the total volume of material available in digital form, attempting to do for digital culture what government, university, museums, and other organized archives have done for print culture — preserve records of the past so as to allow others to access it in the future, including selection, arrangement, conservation, cataloguing — is a major challenge. Even the term "archive" may suggest misleading parallels between older archival practices and what is possible or appropriate for digital materials. People deposit books, personal papers, or theses in an archive, where they remain, unchanged, unless a medium like acid paper demands conservation, for future generations to consult. That may be possible for some resources such as collections of static web pages, as recorded by the Internet Archive, but for dynamic digital resources such as the Orlando Project, archiving even a substantial set of web pages would be only the tip of the iceberg. The Orlando Project has committed to archiving with the University of Alberta Library, which currently entails a fairly well-defined set of practices designed to ensure long-term survival of the data. However, current practices are unlikely to be able to document either the dynamic text or the research process, which was an experiment in large-scale humanities research and computing. We can archive our internal materials, such as meeting minutes, policy documents, and so on. We have an archive of all past versions of the documents that make up the textbase, and of past versions of the delivery system (code and content), so that what it looked like in the past is recoverable. But particular versions will only be recoverable as we have machines that run the browsers and the coding behind them. We need as a community to grapple further with the question of how to archive dynamic resources. 22

Funding bodies such as SSHRC have policies requiring the public archiving of data, even though many researchers are unaware of this requirement and despite the fact that the country lacks the standards and indeed the facilities to permanently archive digital material [MacDonald 2007]. The notice one frequently encounters accessing materials through the Library and Archives Canada Electronic Collection is a sobering reminder of what may be lost: "You are viewing a document archived by Library and Archives Canada. Please note, information may be out of date and some functionality lost" [Disclaimer]. We anticipate out-of-date information in an archive, of course, but if considerable functionality is lost, a digital artifact can hardly be said to have been archived successfully. And this site is devoted to archiving just "monographs and periodicals" rather than more complex artifacts. Archiving an experimental digital project must include the daunting task of somehow preserving not just text but code and functionality, either by maintaining systems on which they can run or migrating them to newer ones. If not, the project will be not done but done for. 23

Evading the Archive

But archiving alone would represent a form of doneness that many digital projects hardly seek. We want *Orlando* to be up and running, to be alive and evolving, being updated and used far into the future. Such longevity in more than an archived state has major implications in terms of resources. Lack of people, time, or funding has consigned more than one project involuntarily to becoming a static tribute to its former activity. The reasons for this include people moving on, intellectually or institutionally, without taking their projects along with them, or people using electronic media to disseminate without particularly desiring to exploit their potential for continual updating, but even where the will to continue persists, inadequate funding mechanisms for sustainability may make it impossible. This is a shame, since, as we have argued here, in the case of *Orlando* and many other digital publications not only does there remain the potential to enrich the contents, but the first iteration often merely begins to tap the potential of the project's data architecture and potential for interface development. Yet once a project publishes a first major release of materials, the assumption that the research is finished makes attracting funding more difficult. While experiments with various models of maintenance and sustainability proceed with both subscription-based and open-access digital publications, it is clear that a fundamental shift is needed in the understanding of the value of project sustainability and ongoing development, along with a concomitant shift of funding models [Our Cultural Commonwealth 2006, 28, 32]. 24

Nor should sustainability be narrowly conceived. Informal user feedback on *Orlando* suggests that, at least with respect to encyclopedic resources, users now firmly expect that scholarly digital publications will be kept up-to-the-minute and respond to user suggestions. Most digital humanities projects presumably would wish to benefit from this respect in which they remain "undone": if we are to evolve useful tools and resources we need carefully to assess how people use them and experiment with ways of making them better. Such inquiry is integral to the Orlando Project's continuing research, since two of its major aims were to establish the viability of extensive, domain-specific semantic markup to enable new kinds of scholarship, and to help shift scholarly users of electronic materials towards more complex engagement with electronic resources. The project also aimed to leverage the markup in ways that we did not have the resources to implement: whether we can ever have done with, that is to have realised, those ambitions will depend on future developments. 25

The Orlando Project directs its research towards two practically inexhaustible fields: women's literary history and the capacity of computing — specifically of extensive XML markup — to serve the needs of this area of humanities inquiry. "Done" becomes, over the course of such an ongoing and complex digital project, a strategic, continually negotiated marker valuable in a range of ways for defining a specific stage of a process which is not unlike that of Lady Mary Wortley Montagu's solo periodical "To be continued as long as the Author thinks fit, and the Public likes it" [Montagu 1993, 105]. Although *Orlando* diverged radically from the sense of authorship invoked here, Montagu conjures succinctly a dynamic relationship between continued production and reception that is as true to the era of digital production as it was to print culture in the eighteenth century. "Done" for *Orlando* the textbase is only newly open, that is beginning, for our users. For this project, so closely focused on a major deliverable, the post-publication phase has simply intensified the importance of the enquiry that binds our two fields of research: that of the relations between *Orlando* and its users. 26

Although they are by no means all unique to digital publication, the factors outlined here, ranging from project conception and design through modes of textuality and publication to complications in sustainability and archiving, work collectively to complicate what "done" means in the context of digital research. They come of participating in a rapidly transforming context for research and publication in the humanities. Many of these threads are tied together by a common concern that has not been present for major projects that issue in print publication: the question of how to make the results of the research continuingly available to others after the point of initial publication. Whatever "done" means for a particular project, those involved face the challenge of ensuring that it does not paradoxically mean a swift end to scholarly circulation and contribution. While a comparison to the loss of the library at Alexandria in the pre-print era might be a tad hyperbolic, it is sobering to contemplate the waste of knowledge and intellectual effort that would result from the failure of the academic community to resolve the thorny problem of how to sustain access, over the long term, to the results of the first generation of experimental endeavours in the digital humanities if we can't figure out what is to be done. 27

Notes

[1]Even with the relatively straight-forward conversion of analog to digital materials, one large-scale archival project found that the cost of an internet-accessible indexed page was about seven times that of the per-page OCR scanning [Internet Library 1999], and the estimate excludes IT and library infrastructure costs. For more complex, experimental, or born digital resources, the delivery development work and costs are bound to be considerably greater.

Works Cited

Brown et al. 2006 Brown, Susan, Patricia Clements, Isobel Grundy. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*. Cambridge, UK: Cambridge University Press Online, 2006. http://orlando.cambridge.org/.

Brown et al. 2006a Brown, Susan, Patricia Clements, Renée Elio, Isobel Grundy. "Between Markup and Delivery; or Tomorrow's Electronic Text Today". *Mind Technologies: Humanities Computing and the Canadian Academic Community*, Raymond Siemens and David Moorman, eds. Calgary, AB: University of Calgary Press, 2006. 15-31.

Bryant 2006 Bryant, John. *Herman Melville's "Typee": A Fluid-Text Edition*. Charlottesville, VA: University of Virginia Press, 2006.

Crane Crane, Gregory, ed. *The Perseus Digital Library*. http://www.perseus.tufts.edu/.

Disclaimer "Disclaimer". *Electronic Collection*. Library and Archives Canada. http://epe.lac-bac.gc.ca/003/008/099/003008-disclaimer.html?orig=/100/205/301/ic/cdc/artist/index.html.

Eaves et al. Eaves, Morris, Robert Essick, Joseph Viscomi, eds. *The William Blake Archive* http://www.blakearchive.org/blake/

Harley 2006 Harley, Diane. *Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences*. Berkeley, CA: Center for Studies in Higher Education, University of California, Berkeley, 2006. http://cshe.berkeley.edu/research/digitalresourcestudy/report/digitalresourcestudy_final_report.pdf.

Hayles 2003 Hayles, N. Katherine. "Translating Media: Why We Should Rethink Textuality". *The Yale Journal of Criticism* 16.2 (2003): 263-290.

Internet Library 1999 *Internet Library of Early Journals (January 1996 - August 1998): A Project in the ELib Programme. Final Report*. March 1999. http://www.bodley.ox.ac.uk/ilej/papers/fr1999/.

Levine-Clark 2007 Levine-Clark, Michael. "Electronic Books and the Humanities: A Survey at the University of Denver". *Collection Building* 26.1 (2007): 7-14.

MacDonald 2007 MacDonald, Moira. "Data Storage Policy Can't Be Enforced". *University Affairs*. 4 June 2007. http://www.universityaffairs.ca/data-storage-policy-cant-be-enforced.aspx.

Mandell Mandell, Laura, ed. *The Poetess Archive*. http://unixgen.muohio.edu/~poetess/.

Montagu 1993 Montagu, Lady Mary Wortley. *Essays and Poems and Simplicity, A Comedy*. Robert Halsband and Isobel Grundy, eds. 1977. Oxford, UK and New York, NY: Oxford University Press, 1993.

Nora *The Nora Project*. http://www.noraproject.org/.

Our Cultural Commonwealth 2006 *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. American Council of Learned Societies, 2006. http://www.acls.org/cyberinfrastructure/acls.ci.report.pdf.

Oxford Dictionary of National Biography 2004-2007 *Oxford Dictionary of National Biography*. H. C. G. Matthew and Brian Harrison, eds. Oxford, UK: Oxford University Press, 2004-2007. http://www.oxforddnb.com/.

Oxford English Dictionary 2007 *Oxford English Dictionary, 3rd ed.*. John Simpson, ed. Oxford, UK: Oxford University Press, 2007. http://www.oed.com/

Tenopir 2003 Tenopir, Carol. *Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies*. Washington, DC: Council on Library and Information Resources, 2003. http://www.clir.org/pubs/reports/pub120/pub120.pdf.

Warwick 2004 Warwick, Claire. "Print Scholarship and Digital Resources". *A Companion to Digital Humanities*. Susan Schreibman, Ray Siemens, and John Unsworth, eds. Oxford, UK: Blackwell, 2004. 366-382.

Warwick et al. 2006 Warwick, Claire, Melissa Terras, Paul Huntington, Nikoleta Pappa, and Isabel Galina. *The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities Final Report to the Arts and Humanities Research Council*. London, UK: School of Library, Archive and Information Studies, University College London. 2006. http://www.ucl.ac.uk/slais/claire-warwick/publications/LAIRAHreport.pdf.

Wiberley 1991 Wiberley, Stephen E., Jr.. "Habits and Humanists: Scholarly Behavior and New Information Technologies". *Library Hi Tech* 9 (1991): 17-21.

Wiberley and Jones 2000 Wiberley, Stephen E., Jr. and William G. Jones. "Time and Technology: A Decade-Long Look at Humanists' Use of Electronic Information Technology". *College and Research Libraries* 61.5 (2000): 421-431

Women Writers Project *Brown University Women Writers Project*. http://www.wwp.brown.edu/.

The Mutual Knowledge Problem and Its Consequences for Dispersed Collaboration

Catherine Durnell Cramton
School of Management, George Mason University, Mail Stop 5F5, Fairfax, Virginia 22030–4444
ccramton@som.gmu.edu

Abstract

This paper proposes that maintaining "mutual knowledge" is a central problem of geographically dispersed collaboration and traces the consequences of failure to do so. It presents a model of these processes which is grounded in study of thirteen geographically dispersed teams. Five types of problems constituting failures of mutual knowledge are identified: failure to communicate and retain contextual information, unevenly distributed information, difficulty communicating and understanding the salience of information, differences in speed of access to information, and difficulty interpreting the meaning of silence. The frequency of occurrence and severity of each problem in the teams are analyzed. Attribution theory, the concept of cognitive load, and feedback dynamics are harnessed to explain how dispersed partners are likely to interpret failures of mutual knowledge and the consequences of these interpretations for the integrity of the effort. In particular, it is suggested that unrecognized differences in the situations, contexts, and constraints of dispersed collaborators constitute "hidden profiles" that can increase the likelihood of dispositional rather than situational attribution, with consequences for cohesion and learning. Moderators and accelerators of these dynamics are identified, and implications for both dispersed and collocated collaboration are discussed.

(*Dispersed Collaboration*; *Dispersed Teams*; *Distributed Work*; *Virtual Teams*; *Mutual Knowledge*; *Information Exchange*; *Information Sharing*; *Shared Understanding*; *Attribution*; *Proximity*; *Computer-Mediated Communication*; *Systems Dynamics*; *Cognitive Load*)

The organization of group work and the means of communication to support it are changing. Developments in communication and collaborative technologies have made it feasible for groups to work together despite physical dispersion of members. Organizations have been quick to experiment with geographically dispersed work teams to take advantage of interorganizational and international opportunities and maximize the use of scarce resources. This is likely to be an increasingly prevalent and important form of work in the years ahead (Arthur and Rousseau 1996, Boudreau et al. 1998, DeSanctis and Poole 1997, Handy 1995, Kemske 1998, O'Hara-Devereaux and Johansen 1994, Townsend et al. 1998).

Geographically dispersed teams are groups of people with a common purpose who carry out interdependent tasks across locations and time, using technology to communicate much more than they use face-to-face meetings (adapted from Lipnack and Stamps 1997, and Maznevski and Chudoba 2000). The use of such teams has outpaced our understanding of their dynamics, and inexplicable problems have been noted. In a field description of dispersed collaboration, Armstrong and Cole (1995, p. 187) observe these puzzles: "A decision made in one country elicits an unexpected reaction from team members in another country . . . Conflicts escalate strangely between distributed groups, resisting reason. Group members at sites separated by even a few kilometers begin to talk in the language of 'us and them'."

This paper utilizes the communications literature on "mutual knowledge" to explore challenges of communication and collaboration under dispersed and technology-mediated conditions. Mutual knowledge is knowledge that the communicating parties share in common and know they share (Krauss and Fussell 1990). In the work of communication theorist Herbert Clark and his associates, mutual knowledge is referred to more broadly as "common ground," and considered integral to the coordination of actions (Clark 1996, Clark and Carlson 1982, Clark and Marshall 1981). But members of dispersed teams do not stand on common ground. Indeed, the usage "common ground" suggests how deeply engrained physical copresence and shared physical setting may be to establishing shared understanding and affiliation. In 1990, Krauss and Fussell raised the question of how the use of new communications technologies to support cooperative work would interact with the problem of establishing mutual knowledge. This paper takes up that question and adds to it two additional ones: "How does geographic

1047-7039/01/1203/0346/$05.00
1526-5455 electronic ISSN

dispersion of team members affect the mutual knowledge problem?" and "To the extent that geographic dispersion and use of new communications technologies affect the mutual knowledge problem, what are the consequences for collaboration?"

The Mutual Knowledge Problem

Establishing mutual knowledge is important because it increases the likelihood that communication will be understood (Clark 1996, Clark and Carlson 1982, Clark and Marshall 1981, Fussell and Krauss 1992, Krauss and Fussell 1990). People may start with the same information, have a shared experience, or share information through communication. In each case, mutual knowledge consists not only of the information itself but also the awareness that the other knows it. For example, Clark (1996) describes standing on a beach on a beautiful day, examining a rare conch shell. If his son joins him, their mutual knowledge now includes the characteristics of the beautiful day, the beach and the sea, the presence of Clark, the presence of the son, the conch shell between them, and their awareness that they share this knowledge in common. When they talk with each other then and later, they can refer to aspects of this experience with considerable confidence that what they say will be understood by the other. They also can coordinate future actions with the help of this mutual knowledge. For example, they could agree to meet back at the same place in an hour.

Mutual knowledge increases the likelihood of comprehension because it allows speakers "to formulate their contributions with an awareness of what their addressee does and does not know" (Krauss and Fussell 1990, p. 112). Proceeding without mutual knowledge, people may speak and understand what is said on the basis of their own information and interpretation of the situation, falsely assuming that the other speaks and understands on the basis of that same information and interpretation (Blakar 1985). Krauss and Fussell (1990) describe three mechanisms by which mutual knowledge is established: direct knowledge, interactional dynamics, and category membership. The next sections draw on several research literatures to consider how dispersed collaboration and use of new communication technologies impact these three mechanisms.

Direct Knowledge

Direct knowledge is created in the course of firsthand experience with individuals (Krauss and Fussell 1990). One can make informed guesses about what they know and do not know on the basis of experiences shared with them and knowledge of their habits and environment gleaned from firsthand observation. The mutual knowledge that Clark and his son have concerning their day at the beach is direct knowledge because it is based on shared experience in a particular setting. For members of dispersed collaborations, opportunities to achieve unmediated knowledge of their partners and their partners' situations are likely to be limited. Rather than absorbing direct knowledge by visiting each other's offices, walking through the same building, attending the same meetings, and driving the same streets, dispersed collaborators must find other means to establish what their remote partners do and do not know: interaction and category membership.

Interactional Dynamics

In lieu of direct knowledge, mutual knowledge can be established through interaction. However, research concerning information sharing and media effects in groups raises questions about the likelihood of success under dispersed and technology-mediated conditions. It is well established that groups that meet face-to-face tend to dwell on commonly held information in their discussions and overlook uniquely held information (Stasser and Stewart 1992; Stasser et al. 1995; Stasser and Titus 1985, 1987). According to Stasser and his colleagues, group members engage in discussion by sampling from their pool of information. As the number of people who have a particular piece of information increases, so does the mathematical probability that it will be mentioned in the group's discussion. Moreover, when it is mentioned, it probably will be salient to a larger proportion of the group because they have encountered it previously. By contrast, pieces of information known to only one or a few people must compete in the information pool with more commonly held information. If mentioned, uniquely held information may not be as salient to group members as commonly held information, and fail to draw attention.

When a group's discussion is mediated by technology, the problem seems to be worse. Three experimental studies compared information exchange in groups using synchronous text-based computer conferencing and face-to-face groups. They found information exchange to be less complete and discussion more biased in the groups using technology to communicate (Hightower and Sayeed 1995, 1996; Hollingshead 1996). The computer-mediated groups exchanged less information overall and took more time doing it. One of the most robust findings concerning the effect of computer mediation on communication is that it proceeds at a slower rate than does face-to-face (Lebie et al. 1996, Straus 1997, Straus and McGrath 1994, Walther and Burgoon 1992). There is considerable evidence that groups using this medium take longer to complete tasks

than groups working face-to-face (Kiesler et al. 1985, McGuire et al. 1987, Weisband 1992). The slower rate has been attributed to the time required to type words rather than say them (Siegel et al. 1986) and the effort required to convey nuances in text without paraverbal and nonverbal cues such as tone of voice, facial expression, and gesture (Hightower and Sayeed 1995, 1996; McGrath and Hollingshead 1994). For these reasons, it appears that groups communicating through such means are not able to sample as much information from their information pool during a given period of time as can groups working face-to-face. As a consequence, less uniquely held information is aired, and their discussion is more biased by commonly held information.

Warkentin et al. (1997) got somewhat different results in a follow-up study under different conditions. They compared information exchange in student groups meeting face-to-face without computer support with information exchange in groups whose members were located at three different universities across the United States and who used asynchronous computer conferencing to communicate. The dispersed groups still exchanged less unique information than the collocated groups; however, the relationship did not reach significance in this study. It is possible that the technology used (an asynchronous rather than synchronous mode) and the longer time frame given the dispersed groups (three weeks versus 25 minutes) allowed for improved information exchange by giving group members offline time to assess information and frame their contributions.

It is important to note how the study conditions differ from geographically dispersed collaboration in practice in organizations. Tasks were relatively straightforward. People did not have to gather information themselves; they were given a packet of clues. By contrast, dispersed collaborators in organizations typically work on complex tasks for which the relevant information must be distinguished from millions of other details. Hightower and Sayeed (1995) note that when groups using computer-mediated communication were given a higher information load, their discussion became even more biased. They express concern about how groups communicating through such media will handle ambiguous and complex information problems.

Dispersed collaboration is distinguished both by heavy use of mediated communication and distribution of partners across more than one location. Thus far, we have examined media effects on information sharing. It is also important to consider the effects on information sharing of differences among locations and distribution of information across locations. The locations of dispersed collaborators may differ in their physical layout and travel requirements, holidays and customs, access to information, available equipment and support, strength of competing demands, and so forth. For dispersed team members to understand each other and coordinate their work, they must achieve mutual knowledge concerning such differences. However, information about one's own location and context may be uniquely held information. According to the principle of group discussion based on sampling from the information pool, such information is less likely to be mentioned and heeded in group discussions than is commonly held information.

When people attempt to achieve mutual knowledge through interaction, they must both share information and confirm that the information has been received and understood. Electronic mediation poses hurdles not only to information sharing but also confirmation. Give and take is hampered by the slower pace and greater effort required by most forms of mediated communication. In particular, conversations conducted through computers do not typically provide efficient back-channel feedback (Brennan 1998). Back-channel feedback includes head nods, brief verbalizations such as "yeah" and "m-hmmm," smiles, and the like (Kraut et al. 1982, Yngve 1970). These nonverbal and brief verbal cues efficiently signal the state of mutual knowledge without taking over the speaking turn. In addition, feedback lags associated with mediated communication and dispersed collaboration are likely to have a devastating impact on the establishment of mutual knowledge. "A delay of 1.6 seconds is sufficient to disrupt the ability of the sender to refer efficiently to the . . . stimuli, despite the fact that the back-channel response is eventually transmitted," observe Krauss and Fussell (1990, p. 132).

Thus, the communication literature raises the following questions: Will uniquely held information be shared and recognized in geographically dispersed, computer-mediated teams? Will feedback in such teams confirm receipt and understanding of information exchanged? In other words, can interaction effectively establish mutual knowledge in dispersed computer-mediated teams?

Category Membership

Finally, people make assumptions about others' knowledge on the basis of the social categorizations they apply to them (Clark and Marshall 1981, Krauss and Fussell 1990). For example, they assume that a cabdriver knows the route to the airport and that a fellow American knows the words to the national anthem. In dispersed collaborations, people may assume mutual knowledge on the basis of shared professional status or organizational membership. However, achieving mutual knowledge will be

more difficult if a dispersed team spans functional, cultural and organizational boundaries. Many dispersed collaborations do span such boundaries.

In addition to influencing assumptions about shared knowledge, social categorization affects the development of relationships among people communicating via computer mediation. Lea and Spears (1991, 1992, 1993) have observed that the medium reduces the number of cues available to communicators about each other relative to face-to-face communication. According to their social identity/deindividuation (SIDE) theory, when people communicate with others they do not know well through such media, they experience feelings of isolation, anonymity, and deindividuation. As a consequence, they "overattribute" on the basis of the few social cues they glean. They use relatively meager information to assign remote others to social categories and treat them accordingly. If they conclude that they share an identity with a remote partner, they are more likely to view the relationship and the remote other positively than if they find no shared social identity. Thus, SIDE theory asserts that social categorizations exert considerable influence over conclusions drawn about remote others, positive and negative feelings about them, and affiliation with them in lieu of the individuating information available in relationships carried out face-to-face.

The salience of social categorization for dispersed collaborators raises the following questions: Does social categorization help dispersed collaborators establish mutual knowledge of important matters? If people assign remote others to social categories on the basis of meager information, will they form inappropriate expectations about what they know? If they form inappropriate expectations, what are the consequences for the relationship?

Consequences of Failure to Establish Mutual Knowledge

The preceding discussion suggests that it may be difficult for geographically dispersed, technology-mediated teams to achieve mutual knowledge. This section considers the consequences of failure to establish mutual knowledge. In the communications literature, mutual knowledge is considered to be a precondition for effective communication and the performance of cooperative work. However, according to Krauss and Fussell (1990), this does not mean that communication must be error free. It depends on whether the consequences of misunderstandings are major or minor, and whether the dynamics of conversation provide mechanisms for detecting and correcting errors. Beyond this, there is little discussion in this literature of specific consequences of failure to establish mutual knowledge. "What is important for us," writes Clark (1996, p. 121) is "how common ground gets staked out and exploited." Therefore, I turn to other literatures to explore the consequences of failure to establish mutual knowledge on decision quality, productivity, and relationships in dispersed teams, taking into account the moderating factors noted by Krauss and Fussell (1990).

Decision Quality and Productivity. As discussed above, when task-relevant information is distributed among members of a group, there is a risk that they will fail to share and heed uniquely held relevant information. Research shows that the consequence is poorer decision quality (Dennis 1996, Stasser and Titus 1985). The risk may be greater for dispersed teams relative to collocated teams for two reasons. First, there is a significant probability that task-relevant information will be distributed across locations and that critical pieces of information will surface at isolated locations. Second, a dispersed group's means of communication will likely restrict interaction, such that sampling from the information pool is less than it would be for a collocated team, with the consequence of poor decision quality.

Groups may attempt to protect decision quality by monitoring and correcting failures of mutual knowledge. In this case, a critical consideration is the ease and speed of detecting and correcting errors (Krauss and Fussell 1990). The more arduous and time-consuming this process is, the greater the probable loss in productivity. This could be a problem for dispersed groups using computer-mediated communication because the interaction required to detect and correct communication errors seems to be less efficient in conversations involving computers than in face-to-face conversations (Brennan 1998; Hightower and Sayeed 1995, 1996; McGrath and Hollingshead 1994; Siegel et al. 1986). Hightower and Sayeed (1995, p. 43) point out that "creating mutual understanding . . . require(s) group members (using computer-mediated communication) to transmit much more information than those working face to face." It is reasonable to expect that dispersed groups using such media will have to sacrifice speed if they must protect decision quality, and decision quality hinges on mutual knowledge of distributed information.

This discussion raises the following questions: When task-relevant information is distributed, do dispersed teams make poorer quality decisions because of failure to establish mutual knowledge? Do dispersed teams develop strategies to limit information-processing burdens and preserve decision quality? Must they accept reduced productivity or are there other options?

Relationships. The work of communication theorist Rolv Blakar and his colleagues provides a critical link

between the concept of mutual knowledge and the consequences for working relationships of failure to establish mutual knowledge. This stream of research investigates how family members react to communication difficulties caused by lack of mutual knowledge. Blakar calls the problem an absence of "shared social reality," however, its operationalization is consistent with lack of mutual knowledge (i.e., people who are engaged in communication have different information but do not realize this is the case). In the studies, pairs of family members are given maps of a city. One subject's map contains arrows that mark a route through the city. This subject is told to describe the route to his or her partner so that the partner can follow the route on his or her own map. Unbeknownst to the subjects, their maps differ in key respects, making it impossible for them to carry out the task successfully. Blakar and his associates were interested in when and how such problems are solved, allowing communication to be reestablished (Blakar 1973, 1984; Hultberg et al. 1980).

They concluded that it is essential that communication difficulties are "adequately attributed" by the participants. When an error or conflict in information exchange is detected, people make attributions concerning its cause. The research focuses on whether attributions concerning communication difficulties are personal or situational (Heider 1958) and constructive or nonconstructive for continued communication. Personal attributions associate the cause of the communication conflict with some characteristic or behavior of an individual. For example, participants in the study conducted by Hultberg et al. (1980) made personal attributions when they made statements such as "My explanation was not adequate" or "You give damned bad explanations!" They made situational attributions when they investigated the credibility of the maps they were using. Attributions were judged to be constructive if they facilitated inquiry and change to reduce the incidence of communication conflicts in the future. Attributions were nonconstructive if they were task irrelevant or destructive to cooperation, inquiry, and adaptation. The researchers suggested that situational as opposed to personal attributions tend to produce better resolution of conflicts because they focus participants on modifying the "contracts" that guide the communication process (Blakar 1984). If attributions are destructive, contracts concerning the communication process break down and people withdraw from cooperation.

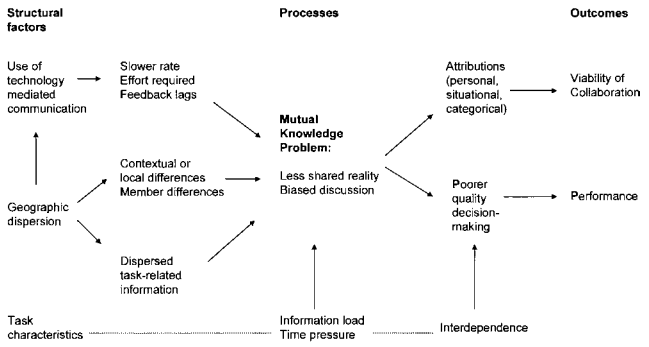
This work has significant implications for the study of mutual knowledge in dispersed collaboration. It recognizes that communication failures are interpreted and that interpretations can vary. These interpretations can change people's perceptions of each other, their willingness to cooperate, and the ways in which they communicate and cooperate. The first part of this section suggests that dispersed collaborations are vulnerable to failures of mutual knowledge. Therefore, the way such failures are interpreted—i.e., attributed—could be critical for the long-term viability of dispersed collaboration.

Blakar and his collaborators focus on the distinction between personal and situational attributions. Applying these ideas to dispersed collaboration, we should consider how the dynamics of social categorization influence attributions and outcomes. As noted previously, Lea and Spears (1991, 1992, 1993) have observed that people using computer-mediated communication with remote others they do not know well rely heavily on social categorizations to guide their relationships. The social categorizations provide a basis for affiliation if participants share a significant social identity. However, they also can provide fodder for in-group/out-group dynamics if remote others are seen as belonging to social categories different and less attractive than oneself. This raises the question of whether or under what circumstances the attributions remote collaborators make concerning failures of mutual knowledge will be personal, categorical, or situational. It is not clear how these different types of attributions affect the viability of dispersed collaboration. Lea and Spears (1992) observe that people tend to overlook errors made by others with whom they share a significant social identity. On the other hand, collaborators might recategorize less generously remote others with whom they experience a communication failure. In addition, we can think of Blakar's "contracts" as norms: understandings within a group about what behaviors are and are not appropriate (Jackson 1965). Ideally, groups that experience communication failures will modify their norms to prevent future occurrences. However, if communication failures are blamed on individuals or subgroups, cooperative norms may break down.

This discussion presents the following questions: How are failures of mutual knowledge attributed in dispersed collaborations? Under what circumstances are attributions individual, categorical, or situational in nature, and with what consequences for future communication and collaboration?

Figure 1 summarizes the relationships discussed in this section. The discussion raises important questions about how mutual knowledge is established and its significance in dispersed collaboration. The existing literature provides a basis for expecting a number of problems with serious consequences. There is a need for close examination of the dynamics of actual dispersed collaborations to see if such problems do occur and how they manifest

Figure 1 **Likely Impact of Dispersion and Mediated Communication on Mutual Knowledge and on Collaborative Outcomes**



themselves. This suggests a fine-grained qualitative analysis of the communication and experiences of groups involved in such collaborations. This approach should illuminate the dynamics suggested by these theories and how they affect participants in ways other types of investigations would not.

Method

The data were contained in an archival dataset that was created in the course of a collaborative project involving graduate business faculty and students located at nine universities on three continents. The project was intended to improve students' technical skills, give them experience in using technology to collaborate with remote partners, and expose them to the possibilities of electronic commerce. I gave my graduate students the option of participating in this project in lieu of the group project I usually assign in my Organizational Behavior course at George Mason University (GMU). Half my students decided to participate in the dispersed teams project. Each of them chose a partner within my class and these pairs were assigned randomly to a team with pairs from two other universities.

The six-member teams included pairs from two U.S. universities and one university located elsewhere. Non-U.S. partners were located in Canada (Université Laval, Memorial University of Newfoundland), Australia (Southern Cross University), and Portugal (Universidade do Minho). There were 45 teams in all, including 13 teams with GMU students. Only the data generated by the 13 teams that had GMU members are used in the analysis, yielding 13 cases. The students ranged in age from 23 to 48 years old, and all were graduate students taking business or management information systems courses. Because a number of foreign exchange students participated, countries of origin included the United States, Canada, Colombia, Portugal, Germany, Ukraine, India, Thailand, Hong Kong, and Indonesia. All of the participants based in Australia were Asian exchange students.

The teams were given the assignment of (1) coming up with an idea for a business that would use the Internet in some way, (2) writing a business plan, and (3) creating a presentation for investors or an online storefront. The project spanned a seven-week period. Communication tools used by the teams included electronic mail, Internet-based "chat" tools, an Internet-based voting tool (Dennis et al. 1996), telephone, and fax. The project home page was a common point of reference for the teams. Home page material included detailed team assignments for each week, sample business plans, project evaluation criteria, and links to resources such as the voting tool, chat rooms, and information about electronic commerce. The teams were permitted to use whatever communication tools they found useful, including those that they found independently on the World Wide Web.

The nine faculty members whose students participated in the project communicated with each other through electronic mail, a faculty listserv, and occasional telephone calls. None of the faculty had met any of the others face-to-face. Thus, the faculty also was a dispersed team that collaborated for three months to manage a complex global social system. Despite efforts to make project requirements consistent across all nine universities, differences were discovered as the project unfolded. When any such difference was discovered, affected students were informed immediately and every effort was made to bring requirements into alignment.

Data Sources

Data constituting the cases include 1,649 pieces of e-mail exchanged by members of the 13 teams, printouts of their online chats, team logs of their use of communication tools, 26 analysis papers written by project participants, and grades awarded independently to each team by two instructors. George Mason University students who participated in the project turned in copies of their e-mail on disk and on paper. On every team on which GMU students worked, there were two Texas Christian University (TCU) students, as well as two students from a university abroad. The TCU students turned in copies of their e-mail, and their professor sent copies to me. I compared the two sets of records and added to the master record newly discovered e-mails. This strategy was intended to identify all cross-site e-mails. As a result of the cross-check, there were only a handful of points in the team histories at which it was clear from the content of the

messages that an e-mail was missing, and I earmarked these points. The count of e-mail records by team ranged from 61 messages (Team 30) to 217 messages (Team 6).

Each team's communication log listed the team's online chats and use of other Web-based tools such as the voting tool. The number of chats held by teams or subsets of teams ranged from none (Teams 11, 18, and 26) to five (Teams 5, 21, and 39). The teams turned in copies of the texts of most of their chats. However, on some occasions they failed to make a printout so the text was not available for analysis. Descriptive information about each of the cases appears on the left side of Table 1.

GMU students wrote individual six-page analysis papers after the project ended. They were instructed to analyze one or two events in the life of their team that they considered to be significant, using e-mail and chat records as a resource. They were required to do their best to examine these events from the perspectives of the other team members as well as from their personal perspective.

To understand the experiences of the teams, I also drew on my experiences as a member of the geographically dispersed faculty team and my work with the students engaged in the project. In addition, several members of the faculty team wrote about their experiences and circulated these narratives. Finally, my colleague at TCU sent me copies of the grades she awarded to each of the 13 teams that included GMU students so that I would have two perspectives, hers and my own, on team performance.

Cursory review of the cases suggested that there had been a great deal of conflict in the teams. In seven of the 13 teams, conflict escalated to the point that hostile coalitions formed. In five of these teams, members at two sites began to complain about partners at the third site, refusing in some cases to send them pieces of the team's work or put their names on finished work. Two teams evidenced shifting coalitions among subgroups at the three sites. Close examination of episodes of conflict, frustration, or confusion in the teams seemed to be merited.

Data Analysis

Data management and analysis procedures are summarized in Table 2. My objectives were to analyze episodes of conflict, frustration, or confusion in the teams, examine the significance and consequences of these episodes in the context of each case as a whole, and look for patterns across cases. I followed Eisenhardt's (1989) specifications for analysis of multiple case studies, with the addition of an embedded information-processing analysis of episodes of conflict, frustration, or confusion within cases (Yin 1994, Coulam and Smith 1985). Steps included (1) putting the 13 cases into an accessible form without compromising their richness, (2) understanding each case on its own terms before attempting to generalize across cases (Eisenhardt 1989, Miles and Huberman 1994), (3) conducting an embedded information-processing analysis of episodes of conflict, frustration, or confusion, (4) creating and refining constructs that cut across cases, (5) identifying other variables of interest, (6) reviewing all cases to refine the definition of constructs and build evidence to measure or refute them, (7) integrating the constructs into a tentative model, and (8) reviewing all cases to refute or refine the model.

Data Management. I followed the process used by Gersick (1988) in her study of eight collocated project teams to gradually condense the voluminous case histories so they could be reviewed telescopically as well as microscopically. I wanted to make it possible to follow the flow of each case while preserving tight links to the original pieces of data. Each of the 13 teams' e-mail was read into AskSam, a text-management software program. Missing e-mail identified through the crosscheck with TCU was added to the files. Each piece of e-mail was assigned an identification number.

AskSam was used to create fields within which to annotate each piece of e-mail. In one such field, my research assistant summarized the literal content of the e-mail. This paralleled Gersick's literal summaries of the team meetings she studied. In another field, my research assistant recorded her interpretations of the activity in the team and the questions that came to her mind. She had been a member of one of the teams and recognized nuances of situations that a newcomer to the complex project probably would have failed to grasp. Her attention was directed primarily to the microscopic level—recording the literal content of each piece of e-mail—with secondary attention to the flow of events.

Case Analysis. I studied each case and recorded my observations in a field created for this purpose. Creation of the summaries made it possible for me to review entire cases quickly when I wished to, tracking the overall flow of events. In addition, counts of the number of e-mails each team exchanged during each day of the project were generated through the software and transformed into graphs of the team's communication activity—another perspective on the flow of the whole. I also examined the e-mail microscopically, comparing my impressions with the summaries and comments of my research assistant. I wanted to be sure that the summaries she prepared were sufficiently descriptive, so that I could rely on them when I wished to move quickly through the material. Using AskSam, one can double-click on the summary of a piece

Table 1 Information by Case

Team	Non U.S. location	Grades	Number of emails	Chats*	Lack of contextual information	Unevenly distributed information	Differences in salience of information	Relative speed of access	Meaning of silence uncertain	Technical problems	Use of external information	Coalition activity
15	Australia	100; 95	156	4	*c*	*u*		*r*	*q*	*T*	*E*	some
26	Canada (Newfoundland)	86; 95**	65	0	*C*	*U*	*s*		*q*		*E*	none
11	Portugal	96; 86	159	0	*C*	*U*	*S*	*R*	*Q*	*T*	*E*	much
30	Canada (Newfoundland)	93; 93	61	3		*U*	*S*		*Q*	*t*	*e*	none
17	Canada (Newfoundland)	90; 90	125	1	*C*	*U*	*s*		*q*	*t*	*E*	some
21	Portugal	90; 90	95	5	*C*	*u*		*r*	*Q*	*T*	*E*	none
1	Portugal	90; 81	85	1	*c*	*U*			*q*	*T*	*e*	some
5	Australia	90; 74	156	5		*U*		*R*	*Q*	*T*	*E*	much
6	Portugal	86; 86	217	1			*S*		*Q*	*T*	*E*	none
3	Canada (Quebec)	86; 86**	153	3	*C*	*u*	*s*	*R*	*Q*	*T*	*e*	much
18	Canada (Quebec)	83; 83	105	0	*C*	*U*	*s*	*r*	*q*	*t*	*E*	none
37	Canada (Newfoundland)	83; 83	130	2	*C*	*U*	*s*	*r*	*q*	*T*	*e*	none
39	Canada (Quebec)	83; 83	142	5	*c*	*U*	*s*	*r*	*q*	*T*	*E*	much

C, U, R, S, Q, and *T* = serious problems of this type in the team. *c, u, r, s, q,* and *t* = some problems of this type in the team.
E = frequent use of information from external sources in the team. *e* = some use of information from external sources in the team.
* represents the number of chats involving at least two locations during which project work was conducted
** number adjusted to reflect differences in grading criteria

of e-mail and be shown the full text. Thus, it was easy to move as I did between microscopic and telescopic views of the team's work. Like Gersick, I eventually condensed the summaries two more times: first into a timeline of key events and then into stages of team activity. I also tracked team behaviors and activities that are generally considered to be predictors of performance and viability: performance strategies, planning activities, leadership activities, roles, coalitions and conflicts, communication with or about outside stakeholders and authorities, and milestones in project development.

After studying a team's e-mail and recording my observations, I reviewed the analysis papers written by the two GMU members of the team. The papers gave me insight into their interpretations of the team's interaction, which I could compare with the e-mail and chat records, my own analysis, and the commentary of my research assistant.

Information-Processing Analysis. Information-processing analyses of collectives such as organizations and teams investigate two central questions: (1) how interactions among members of a collective are influenced by individual information-processing characteristics and limitations, and (2) how structures and systems shape the interactions among individuals and the decisions and actions of the collective. Considerable attention is devoted to examining how individuals "perceive and interpret stimuli and how they remember, use and communicate information about a complex world" (Coulam and Smith 1985, p. 1). Accordingly, information-processing analysis is an appropriate method for investigating the establishment of mutual knowledge and failures of mutual knowledge. The typical medium for information-processing analyses is the case study because of the fine-grained evidence that is required (Coulam and Smith 1985).

I analyzed the exchange and processing of information leading up to, during, and after episodes of conflict, frustration, or confusion in the teams. From e-mail and chat records, I determined what information each team member did and did not have at the time a problem arose. This included activities such as establishing to whom e-mails

Table 2 Data Analysis Procedures

Stages of Work and Key Observations	Activities	Purpose
Data Manipulation		
	Check completeness of e-mail records by comparing e-mail turned in by team members at ECU and SWU.	
	Add to the master file any newly discovered e-mails.	
	Import all e-mail into text management software program.	
	Adjust time stamps to Eastern Standard Time so e-mail can be sequenced correctly and resequence.	
	Assemble records and reports of team "chat" sessions.	
	Research assistant writes a literal summary of each of the 1,649 pieces of e-mail and of each chat session. Focus on content of individual pieces rather than overall flow of events. Observations and questions noted in separate field.	Gersick 1988: Begin process of gradually condensing voluminous transcripts of team activity into summaries of event sequences.
Data analysis		
	Researcher studies each case, moving between summaries and actual text of e-mails. Observations and questions noted in separate field.	Eisenhardt 1989:540; Miles and Huberman 1994: "Allow the unique patterns of each case to emerge" before attempting to generalize across cases."
	Review and write summaries of analysis papers written by members of each team. Papers focused on one or two critical events in the life of the team.	Get perspective on the meaning that participants made of the team's interactions.
	Triangulate researcher impressions with those recorded by participants and research assistant, and with e-mail records.	Develop a deep understanding of each individual case from multiple perspectives.
Observation of high degree of conflict in teams and decision to examine episodes of frustration, conflict or confusion in depth.	Review all team cases and analysis papers to select episodes of frustration, conflict or confusion.	See whether there seem to be some generalizations across cases that could be explored.
	Trace development of each episode, attempting to determine what happened and how each team member saw it.	
Observation that in these episodes, people frequently seemed to be working from different information.	Carefully identify exactly what e-mails and chat experiences the parties to an incident did and did not have at the time it occurred.	
Observation that there seemed to be recurrent types of problems.	Create five constructs representing types of problems observed.	
	Study all 13 cases again in depth to 1) challenge and refine the descriptions of types of problems, 2) document their frequency, 3) document the frequency of other variables of interest, and 4) look for other important types of problems not represented by the constructs.	Eisenhardt 1989: Refine definition of constructs and build evidence that measures the construct in each case.
Observation that there seem to be causal relationships between information exchange activities and failures, attributions about team members, and team coalition activity.	Integrate constructs into a tentative model, drawing on the cases and the relevant literature.	
	Study the 13 cases in their entirety a third time to challenge and refine this model.	Eisenhardt 1989: Verify that the emergent relationships between constructs fit with the evidence in each case.

were and were not addressed, and whether the addresses were correct. I also compared accounts in the student analysis papers with the e-mail and chat records, noting points of divergence in content or tone. In a systematic way, I sought to understand the perspective of each member of a team as an episode unfolded on the basis of the information he or she had at the time and what he or she wrote in e-mail or chats.

Cross-Case Analysis. Across cases, I noticed recurring patterns in the development of problems. I studied these patterns inductively. I characterized individual incidents, grouped them, and adjusted the descriptions and groupings iteratively. Eventually, I characterized five types of problems: (1) failure to communicate and retain contextual information, (2) unevenly distributed information, (3) differences in the salience of information to individuals, (4) relative differences in speed of access to information, and (5) interpretation of the meaning of silence. I reviewed all the cases to refine or refute my descriptions of these problems and assess their frequency across cases and their seriousness within cases. I recorded the identification numbers of e-mails or chats in which I observed each type of problem. Then I reviewed the accumulated evidence in light of my knowledge of each team's case and judged whether each problem was (1) present and serious, (2) present but not serious, or (3) absent. The charting technique is an adaptation of techniques suggested by Miles and Huberman (1994) and used by Elsbach and Sutton (1992). I found I could integrate the five problems into a model that described how episodes typically unfolded and their consequences for teams. In another review of the cases, I sought to refine or refute this model.

Findings

This section reports findings at the episode, case, and cross-case levels of analysis. I describe the five types of problems observed, their frequency and severity across cases, and their association with team strategies and outcomes. I also present a model that summarizes the relationship of problems to each other and the processes observed across cases.

Types of Problems

Failure to Communicate and Retain Contextual Information. Team members had difficulty gathering and remembering information about the contexts within which their distant partners worked. They also failed to communicate important information about their own context and constraints to their remote partners. The teams involved in the project, including the dispersed faculty team, sometimes failed to recognize differences across sites in deadlines for deliverables, evaluation criteria, and the timing of spring break. Only belatedly did many teams come to understand competing time commitments that were affecting members' participation. Team members also sometimes assumed that the collocated partners were in closer touch with each other than they were.

For example, Team 26 experienced conflict over whether or not to schedule online "chat" sessions among team members. The GMU pair refused to schedule a session. According to a GMU partner, we "felt strongly that a phone call would be much more efficient." In his analysis paper, he interpreted the TCU partners' insistence on using the medium as follows: "Perhaps they did not realize how difficult it is to implement the necessary scheduling." What was not communicated in the team was the fact that using chat tools was part of the evaluation criteria at TCU but not at GMU. The TCU students were taking a management information systems course, while the GMU students were taking an organizational behavior course. Despite efforts to make requirements consistent, differences such as this one surfaced. The conflict stemmed from a difference in the *organizational contexts* of the GMU and TCU members, but to the end, the GMU partners saw it only as a *preference* on the part of the TCU partners. Important information about the organizational context and constraints was not communicated among dispersed team members.

In Teams 3, 6, 17, and 18, some team members disappeared during what turned out to be their spring break. One of Team 18's GMU partners believes that she sent her remote partners an e-mail that described GMU's upcoming spring break and stated that she and her collocated partner would be away for three days. When I studied the e-mail records, I could not find her message. At this point, the team had turned in its business plan, but faced an impending deadline for its presentation. After several days passed without e-mail from GMU, exasperated team members at TCU wrote to their partners at Université Laval in Canada: "I can't believe what I just heard. Is it true that Anna and George have spring break now???" Replied the Canadians, "Maybe they are out of town!" Team 3 had a similar experience when partners at Memorial University of Newfoundland who were responsible for assembling all parts of the business plan vanished without warning from the e-mail traffic for three days. Panicked team members at other schools eventually took over the task of their silent partners. The Canadian partners had not mentioned their upcoming break to their distant team members.

In Team 6, e-mail records show that a team member *did* warn the others about a trip during spring break; however, the information did not seem to register in the minds

of the remote partners. Requests from teammates for immediate action continued to arrive in her e-mail while she was away. It appears to be difficult for teammates to create a mental map of their distant partners' situation and to update that map when new information arrives.

At the outset, some students failed to send an initial e-mail message to their teammates for up to three weeks. Often, persons reported later that they were taking midterm examinations or completing some other project. However, they failed to perceive the need to describe their situation to remote partners whose situations might be quite different. Relationships fared better when preoccupied teammates laid out immediately the constraints under which they were operating.

Participants also seemed to exaggerate the completeness of communication in the collocated condition. This is yet another way of misjudging the context of remote partners. A GMU partner in Team 11 assumed that concerns she had expressed on the telephone to one partner at TCU would be conveyed to the other partner at TCU. Later the GMU partner wrote that she had *repeatedly* asked TCU to make a particular change in the team's home page, while one of the TCU teammates insisted that he had *never* heard this concern. One factor in this difference was the GMU partner's assumption about the ease and completeness of communication between the collocated TCU partners.

Unevenly Distributed Information. Unevenly distributed information also interfered with team-level collaboration and caused problems in relationships. Two causes were errors in e-mail addresses and failure to send copies of e-mail to all team members. Team members also may have thought they sent e-mail that in fact never went out or was undelivered. The bucket of information being passed among team members proved to be far leakier than they realized. Wildly different perspectives among team members were created because of differences in the information they received.

In two teams (Teams 37 and 11), partners located outside the United States were quick to initiate communication, yet this failed to become common knowledge in the team. On the contrary, the impression grew that these partners were absent or unwilling to communicate. Meanwhile, they wondered why their initial messages were ignored. In Team 37, a Canadian partner was the second person to check in. After introducing herself, she explained that her collocated teammate was away for a few days but would write shortly. She mistyped the address of one GMU partner, but typed the second GMU address correctly. However, neither GMU partner ever indicated having received her e-mail. E-mail records confirm that at least one of the TCU partners received the e-mail, however this person did not correct the GMU partners when they complained that the Canadians had not been heard from. The dynamic created in the team concerning the Canadians teammates' tardiness persisted to the end, even though it was based on inaccurate information.

In Team 11, the Portuguese partners were the first to write to the group. Their message contained two incorrect e-mail addresses and two correct addresses. It was five days before one GMU team member discovered the message from Portugal in her e-mail. By that time, both GMU and TCU members had begun to worry and complain about what they thought had been silence from Portugal. The Portuguese probably were receiving mail from the United States but wondering why their greeting was being ignored. Moreover, there is no evidence that the GMU partner who found the e-mail from Portugal ever forwarded it to the rest of the team. The Portuguese operated on the assumption that the entire team had the information contained in the initial note, but in fact, only one or two United States members had it.

Impressions formed as a result of unevenly distributed information persisted in the face of correcting information. This is not surprising when one considers how difficult it is to trace all the ways in which a particular piece of information (in this case, erroneous information) has shaped one's feelings about another person. The history of Team 30 offers a striking illustration. The mailing list used by one of the GMU partners, Paul, had an error in it: It included one person who was a member of a different team and omitted one person (Don) who was a member of Paul's team. By the time the situation came to a head, Paul thought he had sent six e-mails to the entire team. He was receiving e-mail from his team member, Don, but did not realize that he was not sending mail to Don. The person who was receiving the e-mail in error never notified Paul. Eventually, Paul spent several hours investigating chat room sites and proposed one to the team. He was astonished when Don wrote within hours proposing a different chat room, assuming that the team would meet there, and ignoring Paul's message. In his analysis paper, Paul wrote that this indicated "resentment toward me for taking the initiative and making decisions."

Paul gave Don the benefit of the doubt by writing to check the e-mail address, but Don did not respond to this note, which offended Paul even more. But let us look at this from Don's point of view: Even though Paul has written six detailed e-mails to the team, Don has not received *any* of them. To Don, Paul is a deadbeat team member who now has finally sent a short note asking if he has Don's address right! Eventually, Paul removed the incorrect name from his mailing list and added Don's name,

but Paul did not change his understanding of the team or see how differently this exchange must have appeared to Don. Even though Paul figured out that he had Don's name and address wrong and offered to send Don all the early messages he had written, Paul still presented the exchange as a power struggle in his analysis paper. Paul still thought that Don deliberately ignored the work he had done.

In relationships conducted face-to-face, it is a challenging cognitive exercise to interpret a set of facts from the perspective of another person. It is far more difficult to determine how the information before the other party differs from one's own, and then see things from the other's perspective. Geographic dispersion makes these two activities more difficult because of undetected "leaks in the bucket," because partners seem to have difficulty retaining information about remote locations, and because feedback processes are laborious. In addition, the data suggest that team members with complete or correct information may not speak up when erroneous conclusions are voiced in the team.

Problems stemming from unevenly distributed information were not limited to cases involving errors in addresses and undelivered mail. Sometimes people knew they were exchanging mail with only part of the team, but failed to understand how this affected the perspectives of team members who did not receive the mail, or how it affected the dynamics of the team as a whole. In Team 11, Lisa in Portugal suggested a focus for the team's project. Team members at GMU and TCU exchanged e-mail about Lisa's idea without copying her. They agreed that her idea was creative and interesting but too complex for the team's time frame. When the team voted electronically, Lisa's idea was not selected. From Lisa's perspective, her idea was met by silence. There was no discussion, praise, or criticism—just a vote. When team members began volunteering via e-mail to develop particular parts of the business plan, the Portuguese part of the team was silent. I suspect that the GMU and TCU members were more conscious of the silence from Portugal than their own silence in response to Lisa's idea. E-mail concerning Lisa's idea was exchanged. It just wasn't sent to Lisa.

It is not news that private conversations can create problems in a team. However, the dynamics and consequences of this behavior in dispersed teams are worth noting. Private exchanges of e-mail distort perceptions of the volume of activity in a team. This can confuse members' sense of pace and timing. Members who are receiving all the e-mail will perceive some members as active and others as relatively inactive. Those who are not receiving all the mail will perceive the energy level of the team to be low and the pace to be slow. Accordingly, they may further reduce their pace, or berate members for inactivity because they do not know of their efforts. Analysis of the team histories suggests that these kinds of perceptions can be excruciatingly difficult to identify and change when a team is dispersed. Private "conversations" may create much more confusion for dispersed teams than face-to-face teams.

Differences in the Salience of Information. Teams also encountered problems that hinged on differences in the salience of information among team members. Writers tended to assume that what was salient to them would be salient to their readers. Scholars have observed that mediated communication often lacks cues to *meaning* such as facial expressions, body language, and tone of voice (Kiesler and Sproull 1992, Sproull and Kiesler 1986). However, these types of cues also signal the *salience* to the communicator of one piece of information relative to another. Dispersed team members were not successful in communicating to their partners what parts of their messages, or which messages, they considered most important. When an e-mail message addressed several topics, partners sometimes differed on which topics they found salient. For example, as described previously, a GMU partner who wanted a change made in the team home page raised the issue on the telephone with TCU Partner 1 and in a postscript to a four-paragraph e-mail to TCU Partner 2. The proposed change was the last of three issues addressed in the e-mail. TCU Partner 2 later insisted indignantly that he had never heard of the request. Clearly, the postscript had greater salience to the sender than to the receiver.

Differences in information salience were exacerbated by unwieldy feedback processes in the dispersed teams and the making of indirect requests. Analyzing a tense exchange in Team 6, a GMU member observed:

> With so much information going back and forth, it was difficult for my teammates to absorb every detail . . . Because I couldn't "see" if the receiver was paying attention, I didn't know if my message had to be repeated. Yet it is time-consuming to let the sender know my perception of their message.

In other cases, writers did not use in their e-mail key words that they thought they had used. In Team 18, Anna wanted to clear up irritation that had resulted when she and George disappeared for three days during GMU's spring break. In her analysis paper, she wrote, "When everything was done, I thought it was time to clear our misunderstandings. I didn't want to brush things away. I sent an e-mail saying that we need to have a chat as there were some misunderstandings to be cleared." The chat never took place. One reason may be that Anna did not

actually use the words "clear our misunderstandings" in her e-mail. Instead, she wrote, "There seems to be a communications lapse between us. George and I thought that we probably need to discuss certain aspects from the home page. Is it possible to chat today? . . . This is important, so please let us know soon." It is unlikely that Anna's wish to resolve the issue of why she and George disappeared was clear to her dispersed teammates because she said that she wanted to "discuss certain aspects from the home page." Confusion due to indirect wording is not confined to computer-mediated communication, as Tannen (1994) has shown. But the characteristics of some of the communication technologies used by dispersed groups probably make it difficult for members to recognize the meaning and importance of indirect requests like this one.

In addition, there was a tendency to request feedback from the team indirectly, yet to expect quick responses from everyone. (Kiesler et al. 1984 predicted the latter.) "Every time I sent an e-mail requesting "any thoughts" from everyone, I expected to receive one from everyone. And when I didn't, I felt that those who didn't respond were not holding up their end of the bargain," reported a member of Team 5. A member of Team 17 observed, "People always said, 'Hope to hear from you soon.' Who then has responsibility for initiating communication?" A member of Team 11 recalled sending e-mail "into the abyss," and GMU-based members of Team 6 discussed the feedback problem among themselves, but never with other members of their team. In an analysis paper, one wrote:

> We wanted acknowledgment of the time we spent on the deliverable as well as a feeling that we were on target. No one responded. We sent another e-mail saying that we hadn't heard from anyone. Finally, we heard from one group member, but even that message contained minimal information.

Clearly, the salience of the request for feedback was higher for senders than receivers. Thus, when electronic communication is voluminous, senders and receivers unwittingly may differ in what they find most salient and fail to fulfill their distant partners' expectations. This problem may be complicated by a tendency to state requests indirectly, yet expect quick responses from all members. In general, the level of feedback among members of a dispersed group is not likely to be as high as members would wish, and may not be sufficient to ensure shared understanding.

Relative Differences in Speed of Access to Information. Research has shown that teams using computer-mediated communication operate at a slower rate than teams meeting face-to-face (Lebie et al. 1996, Straus 1997, Straus and McGrath 1994, Walther and Burgoon 1992). However, this study surfaced a second type of problem involving speed: relative differences among team members in speed of access to information. One manifestation of this problem stemmed from differences among team members in access to communications technology. Some members had 24-hour e-mail access while their partners had access only when at their university. If some members see e-mail only once a day or once every few days, this limits the interaction that is possible and slows the pace of the team. Observed a member of Team 3, "Some problems dragged on for days while the suspicions of group members intensified. In reality, the problem could have been as simple as someone not being able to get to the computer lab to check their messages."

A second manifestation seemed to stem from differences in the speed of electronic transmissions among parts of a team. This was exemplified by the relationship among members of Team 5, which included two members in Australia and four members in the United States, two of them in Virginia and two in Texas. The team held five online chats during which tensions between the members in the United States and Australia were evident. Near the end of the team's fifth chat, an American team member observed that the Australian members always seemed to be "25 minutes behind the discussion" and suggested that this could be an artifact of the speed of transmission between the continents. The team members at the two United States locations could carry on a relatively rapid exchange until being "interrupted" by team members in Australia who referred to subjects from which the others had moved on.

I was not able to verify this hypothesis through chat room transcripts or other means. However, the explanation proposed by the team member is credible. Telephone lines carried most of the Internet traffic between the United States and Australia and they frequently became overloaded, resulting in breakdowns and time lags. This would mean that parts of the team were communicating at different rates—one rate between the two sites in the United States and another rate between the United States and Australia. This is a recipe for frustration and irritation for all. If the members in Australia responded to messages the instant they received them, their responses still would appear in the chat room traffic well after the conversation between the United States partners had moved on because of time lags coming and going. Moreover, from the perspective of the Australian partners, a stream of unrelated comments by the United States partners would always follow their messages. It would appear that their comments were ignored.

Both types of problems concerning relative speed

tended to be invisible to team members. Instead, they were attributed to remote partners' lack of conscientiousness. In addition, recognizing the constraints on access to information was only of limited help to the teams. The fact remained that it was difficult for team members to work "in sync" with one another. Although the source of problems sometimes was identified, suspicions and impressions that had formed tended to persist.

Interpreting the Meaning of Silence. One of the biggest challenges team members faced was interpreting the meaning of their partners' silence. Over the course of the project, it became clear that silence had meant all of the following at one time or another: I agree. I strongly disagree. I am indifferent. I am out of town. I am having technical problems. I don't know how to address this sensitive issue. I am busy with other things. I did not notice your question. I did not realize that you wanted a response.

Partners often misinterpreted silence. One common problem was interpreting silence as consent when it stemmed from disagreement or inattention. For example, United States members of Team 11 misjudged the silence of their Portuguese partners after an electronic vote. The Americans interpreted the silence as consent, and then began to wonder. When they inquired, their Portuguese partners replied, "Yes we are still (here), but you had decided everything. Now you should tell us what you want . . . We don't know (anything) about (the business idea that was chosen)." Similarly, a GMU member of Team 6 informed her teammates that she would be away during spring break. "I asked if I needed to submit anything for the home page before I departed . . . When I didn't receive a response, I assumed everything was in order." After she had left, her teammates began to write terse e-mails, asking for the address of her personal home page, which they wished to link to the team home page.

Silence due to technical problems or faulty information sometimes was interpreted as intentional nonparticipation. A member of Team 21 became concerned when various team members "did not respond to most of the e-mails and kept missing chat room meetings." He interpreted this as his partners' "unwillingness to work." It eventually was discovered that the U.S. partners thought the time difference between themselves and their Portuguese partners was six hours, when it was five hours. Times for the synchronous "chats" among the partners were communicated in error, which meant that the Portuguese partners would arrive at the appointed hour and find no one there. This confusion persisted over the course of four chats and two weeks before it was ironed out.

In meetings conducted face-to-face, it can be difficult to interpret the meaning of team members' silence. However, geographic dispersion and reliance on communications technology add new dimensions of uncertainty and complicate efforts to resolve the uncertainty. A partner could be out of town or silenced by technical problems. There may be a tendency to fall silent rather than address sensitive issues because of the difficulty of communicating nuances when using less rich communications media. In particular, uncertainty about silence can make it difficult to know when a decision has been made in a geographically dispersed group.

Clearly, the 13 dispersed teams struggled with problems involving the distribution and interpretation of information. The problems were serious. They affected individual working relationships and the viability of teams as wholes. Relationship problems were created and magnified by flaws in information management. These problems were difficult to correct. Corrective feedback was scarce and slow, and it was laborious to modify impressions in the face of new information. Impressions persisted in the face of corrected information. Table 3 summarizes the five types of problems identified.

Frequency of Problems

Table 1 displays the occurrence by team of each type of problem described above. The display includes indicators of team activity (the total number of e-mails logged, the number of online chats held, and communication with external sources of information), in-group/out-group dynamics (whether hostile coalitions formed), and performance (grades received). Also noted for each team is the occurrence and severity of technical problems such as the inability to access groupware, chat line, or e-mail servers when planned, the inability to transmit compatible files among team members, and the use of incorrect passwords or procedures. Table 4 summarizes the frequencies of occurrence and severity of each problem across all the teams. It shows that the most severe disruptions to teams' work were caused by uneven exchange of information and technical problems. These problems were serious in nine of the 13 teams. The most common problems across teams were difficulty interpreting the meaning of silence, uneven exchange of information, and technical problems. Difficulty interpreting the meaning of silence was "some problem" or "a serious problem" in all 13 teams. Uneven exchange of information and technical problems were "some problem" or "a serious problem" in 12 of the 13 teams.

Association with Strategies and Outcomes

I looked for relationships between the incidence and severity of particular problems in teams and a measure of

Table 3 Types of Information Problems

Failure to communicate and retain contextual information
Propositions: It is difficult for dispersed collaborators to gather, retain, and update information about the contexts in which their distant partners work, particularly as the number of locations increases. Reciprocally, dispersed collaborators often fail to communicate important information about their own context, situation, and constraints to their remote partners.
Examples of such information include the length of the trip to the office, the quality, accessibility, and features of equipment, measurement processes and standards; local holidays and customary practices, pressure from local supervisors and coworkers, local history and interpretive schemas, competing responsibilities, and local emergencies.
Unevenly distributed information
Propositions: Dispersed collaborators fail more often than they realize to distribute the same information to all members. Causes include human and technological error, and selective distribution without awareness of all its consequences.
Uneven distribution of information results in team members having different perspectives because of the different information they have. Partners with complete or correct information may not speak up when erroneous conclusions are voiced in the team. Impressions created on the basis of unevenly distributed information often persist in the face of correcting information. Uneven distribution of information distorts perceptions of the volume of activity in a team, and confuses the team's pacing and timing.
Differences in the salience of information among members of a dispersed collaboration
Propositions: Dispersed collaborators tend to be less successful than collocated collaborators in communicating to their partners what parts of their messages, or which messages, they consider most important. They may assume that what is salient to them will be salient to remote partners. In particular, requests that are stated indirectly may be salient to the person making the request but not salient to the object of the request. While this problem is not unique to dispersed teams, it may be problematic for them because of restricted back-channel feedback and often slow feedback cycles.
Relative differences in speed of access to information
Propositions: Dispersed collaborations are susceptible to problems that stem from parts of a team communicating at different rates. Some members may be in frequent contact with one another while others are heard from less often. Causes include differences among parts of a team in the speed of electronic transmissions or in access to communications technology. This means that partners are not synchronized in terms of their access to information and their ability to detect and correct misunderstandings. The structural causes of these types of problems tend to be invisible to team members. Even when recognized, differences in communication rate pose problems for collaboration.
Interpretation of the meaning of silence
Propositions: Dispersed collaborators often are uncertain about or misinterpret the meaning of their remote partners' silence. Geographic dispersion and reliance on communications technology add new dimensions of uncertainty to the meaning of silence and complicate efforts to resolve the uncertainty. There may be a tendency to fall silent rather than address sensitive issues because of the difficulty of communicating nuances using the available media. Uncertainty about silence can make it difficult to know when a decision has been made in a geographically dispersed group.

Table 4 Frequency of Problems Across Teams (in percentages)

	Type of problem					
Extent of problem	Lack of contextual information	Unevenly distributed information	Differences in salience of information	Differences in speed of access	Meaning of silence uncertain	Technical problems
Serious problem	54	69	23	23	46	69
Some problem	23	23	46	38	54	23
Not a problem	23	8	31	38	0	8

performance (i.e., grades received), the incidence of hostile coalitions, and the number of e-mails and chats logged. I also looked to see whether teams could be grouped by their task and relationship management strategies. The data suggest four constellations: (1) good performance, task focus, moderate relationship demands, relatively low volume of communication, and low coalition activity; (2) good performance, high task and relationship demands, relatively high volume of communication, and high coalition activity; (3) weaker performance, relatively high volume of communication, many and diverse information problems, and high coalition activity; and (4) weaker performance, relationship focus, task secondary, relatively high volume of communication, and low coalition activity. There were no obvious direct relationships in this small sample between the incidence and severity of particular information problems and performance. Information problems seemed to be more damaging to relationships than to task performance. It also seemed to matter how members dealt with the information problems that occurred: the extent to which they extended the benefit of the doubt when remote partners did not behave as expected, and the extent to which they were able to identify aspects of the situation that helped explain behavior that was contrary to expectations.

Thus, teams with an average grade of 90% and above seem to constitute two types: those that exchanged relatively little e-mail and did not develop coalition activity, and those that exchanged more e-mail and developed coalition activity. Teams 26, 30, and 21 fit the first category and Teams 15, 11, and 17 fit the second. The three teams that earned a grade of 90 or above and avoided the development of hostile coalitions emphasized minimalism and closure in their work processes and tended to give remote partners the benefit of the doubt. In other words, they kept task and social demands relatively low. They exchanged some personal information at the outset, but otherwise tended to focus on the task at hand. The tone was businesslike but gracious. Through skill or luck, they agreed quickly on a viable idea for a business and then focused on implementing the idea. Giving remote partners the benefit of the doubt was important, as each of these teams did encounter significant information exchange problems early in their lives that easily could have escalated. However, members' interpretations of these events, attributions made about remote partners, and consequent behaviors did not accelerate the development of hostile relationships in the team. They struggled to maintain a gracious tone despite unanswered questions, seeming slights, and frustration.

For example, Team 26 exchanged only 65 messages, less than half the number that many other teams exchanged. The average grade it received on its project was second highest among the thirteen teams studied. Individual introductions were brief and businesslike, but the tone of communication was gracious overall. One team member repeatedly modeled this behavior, while still taking a strong stand on a few issues. The team's communication is peppered with comments such as "Good ideas!," "Thanks for the vote of confidence," "Great pic," and "What a talented group!" The group encountered a serious challenge early on because of errors in e-mail addresses involving both the GMU and TCU partners. The upshot was that four of the six team members voted to accept a project idea that the two GMU members had never even received, while a project proposal from GMU bounced back undelivered from TCU. Both GMU members noted in their analysis papers that they were irritated by this event; however, one of them added, "I do understand why this occurred. An initial phase of confusion is to be expected in virtual teams that rely primarily on e-mail for communication." In other words, he made a situational rather than personal or categorical attribution. When new challenges involving differences in context, constraints, and ideas arose for the team, they were handled in sophisticated ways. Confronting disagreements over elements of the business plan, a team member in Canada wrote, "Greetings, friends. We seem to be at odds over what our company should be doing . . ."

In the other three teams that earned a grade of 90 or above (Teams 15, 11, and 17) some members were determined to turn in a quality product and managed to do so. However, life inside the team was turbulent. Attempts were made to extend the benefit of the doubt to remote partners; however, attributions grew harsh. For example, a member of Team 11 eventually wrote, "We got a cheap excuse (from the Portuguese partners) that I did not buy." Ultimately, partners at two of the three locations began planning to do the project as a group of four and discussed whether to withhold the names of the out-group members from the finished product. It may be that the approach to the task these teams took was relatively demanding, and it overwhelmed the team's social capacity and available communication media. For example, these teams debated longer on the kind of business they would design than did Teams 26, 30, and 17. When information problems occurred, they were difficult to resolve and did considerable damage.

The same problems and processes appeared in a more virulent form in Teams 5, 3, and 39, whose grades averaged below 90 percent and who had a relatively high volume of communication, frequent and diverse information problems, and fierce coalition activity. Almost all

of the teams with much coalition activity included students who were high achievers, judging from the quality of their work in my course overall. They may have demanded more from the dispersed mode of work and their remote partners than could be delivered. For example, Team 5 subjected ideas to intense scrutiny but struggled in the aftermath to maintain working relationships. Reflected a member of Team 3, "Our group lacked any real talent for offering compromises and moving on." Another member of Team 3 described the behavior of partners as "belligerent, lackadaisical, and indifferent." These groups seem to have become caught in destructive and self-reinforcing patterns of interaction to which dispersed teams are susceptible.

On the other hand, Teams 6 and 37 privileged harmony over quality. Team 6 is an outlier in that its e-mail volume was the highest of all the teams and its internal relationships appear to have been the most positive. Members of this team realized early on that uneven distribution of information among team members could be a problem and invested considerable effort in avoiding this situation. However, the team's work product was not graded as highly as that of some of the other teams. In analysis papers, members of both Teams 6 and 37 suggested that desire for harmony in the team had interfered with scrutiny of business ideas. Wrote a member of Team 37, "(We) felt we should go along with the other members' ideas . . . even though we did not agree with them . . . The whole team was never in conflict. Everyone was very polite."

Integrative Model

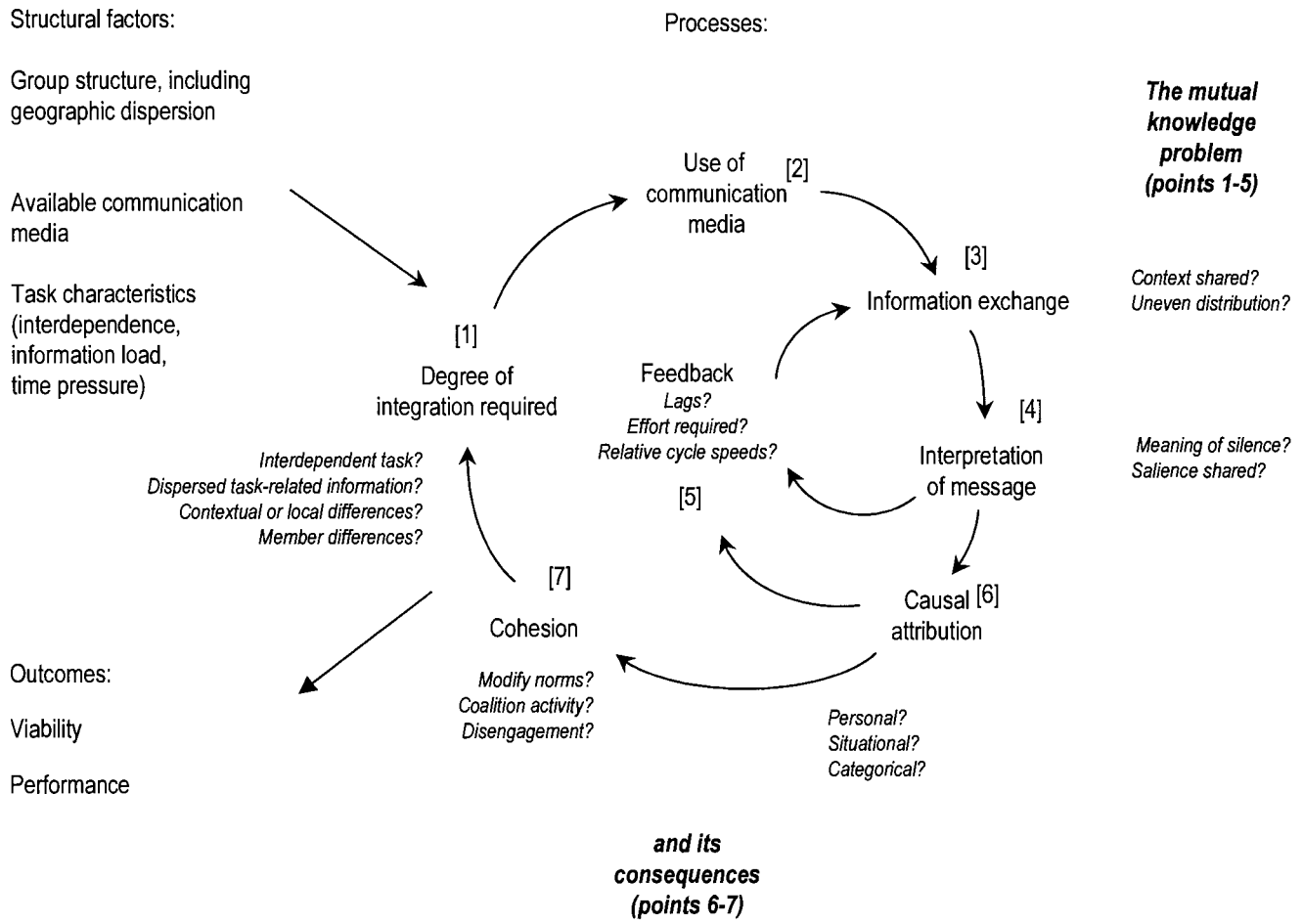
Figure 2 summarizes the processes I observed. It focuses on the challenges dispersed teams face in integrating individual member contributions and maintaining social integration in light of the mutual knowledge problem and its consequences. The task requirements, context, and composition of the group establish the degree of integration required for effective performance and how difficult to achieve this is likely to be (see Point [1] in Figure 2). For example, a complex task may require a high level of integration of individual expertise for success. However, achieving this integration will be more challenging if group members start from different social and practical realities—e.g., come from or live in different cultures, have different functional backgrounds, are working across organizational boundaries, or are widely dispersed geographically from one another. In addition, the characteristics of the available communication technologies, their appropriateness for the task at hand, and how the group uses these technologies (Point [2]) represent another set of enabling and constraining conditions (DeSanctis and Poole 1994, Maznevski and Chudoba 2000, McLeod 1996).

Failure to exchange adequate information about context and failure to distribute the same information to all members of the team constitute two major pitfalls in information exchange (Point [3]) to which dispersed teams may be subject. It may also be difficult for team members to retain and update information about remote contexts provided by their distant teammates. As a consequence of these failures, group members are more likely to work from different definitions of the situation, which handicaps communication and collaboration. In addition, failure to share and retain up-to-date information about context contributes to communication breakdowns by limiting the ability of senders to frame communication in a decentered rather than self-centered way. Messages framed in a decentered or receiver-centered way have a better chance of being interpreted accurately (Point [4]) (Blakar 1985). The likelihood of tension is probably greater if group members' social and practical realities are quite different from the outset (Point [1]); however, subsequent aspects of interaction may add to the confusion. Human error in using technology and technical failures can create information distribution problems in an otherwise untroubled situation. However, once tensions have begun to develop, negative attributions have been made (Point [6]) and coalitions have formed (Point [7]), members may be more inclined to distribute information selectively among team members, and problems escalate.

Exchange of information is just one part of the communication process. Receivers must "decode" the symbols that constitute a message and interpret the meaning of the sender (Redding 1972, Rommetveit 1968). The contexts within which senders and receivers encode and interpret information are likely to differ when their geographic locations are distant, increasing the likelihood of misinterpretation. This problem is exacerbated by failure to exchange sufficient information about context. Two other possible pitfalls at this stage (Point [4]) of the process are the drawing of erroneous conclusions about remote partners' silence, and differences in the salience to sender and receiver of different parts of a message, leading to differing interpretations of the message.

The available information and how it is processed affects attributions (Point [6]). I propose that personal attributions are made about remote partners more often than collocated partners because more information is available about the local than the remote situation. The complexity of dispersed structures and processes also makes situational attribution difficult. Interpretations and attributions can be checked through feedback, however geographically dispersed groups face three challenges: time lags,

Figure 2 The Mutual Knowledge Problem and Its Consequences for Dispersed Collaboration



the effort required to seek and give feedback when dispersed, and relative differences that may exist in feedback speeds among parts of the group (Point [5]). Slow feedback cycles (Point [5]) reduce corrective feedback and increase the likelihood of erroneous interpretations (Point [4]) and exaggerated attributions (Point [6]).

Finally, there may be a tendency to generalize such social perceptions, particularly negative ones, to the locational subgroup of which a person is a member, which sets in motion in-group/out-group dynamics (Point [7]) that are destructive to group cohesion. In some cases that I studied, collocated partners relied increasingly on each other, criticizing their remote partners among themselves and sometimes disengaging from the group's work. In other cases, subgroups in two locations exchanged critical e-mail about the third subgroup and refused to send them work products. Typically, an important piece of information was not sent, sent to the wrong place, or lost in transit (Point [3]), which affected the conclusions drawn by all parties (Point [4]) and led to negative attributions to individuals and subgroups (Point [6]) and the disintegration of relationships (Point [7]). Sometimes, problems were mitigated when errors were caught in the course of a feedback cycle (Point [5]). However, there were instances in which this correction was so slow in coming that those involved were unable to trace and modify all the faulty conclusions they had drawn. The cycle depicted in Figure 2 can be self-reinforcing: The problems of information exchange, interpretation, and attribution described here, and their disintegrative effect on team relationships, add to the already substantial integration challenge confronted by a geographically dispersed group (Point [1]).

Discussion

This paper proposes that a central problem of geographically dispersed collaboration is maintaining mutual

knowledge. Both physical dispersion of collaborators and frequent use of communications technology tend to negatively affect the means by which people establish mutual knowledge. I also suggest that failure to establish and maintain mutual knowledge can have serious consequences for the viability of dispersed collaboration. My empirical findings support and develop this theory. The five specific problems I identified inductively are manifestations of the mutual-knowledge problem that are especially likely under conditions of physical dispersion. I describe ways in which these problems interact, exacerbating factors, and typical consequences. While failure to establish and maintain mutual knowledge may be more likely and serious when collaborators are physically dispersed, the problems, processes, and consequences are probably not limited to dispersed collaboration. I offer grounded suggestions about how and when these dynamics might appear in collocated collaborations.

Failures of Information Exchange

Two specific problems that came to light in this study concern failures of information exchange which result in dispersed partners having different information, but not knowing this is the case. Members of the teams I studied often failed to guess which of the many features of their context and situation differed from the contexts and situations of remote partners. They did not communicate critical local information. Second, team members failed far more often than they realized to distribute the same information to all members. Causes included human and technological error, and selective distribution without apparent awareness of all its consequences.

Identification of these failures of information exchange extends our understanding of both mutual knowledge and dispersed collaboration. One reason that geographic dispersion poses challenges to collaboration is that locations are likely to differ. Differences can include the length of the trip to the office; the quality, accessibility, and features of equipment; measurement processes and standards; local holidays and customary practices; pressure from local supervisors and coworkers; local history and interpretive schemas; competing responsibilities; and local emergencies. In addition, dispersed teams may be more likely than collocated teams to include members with different cultural backgrounds and organizational affiliations, which introduce still more contextual differences. People who wish to collaborate must discover and work across these differences.

However, my research suggests that dispersed collaborators are not skilled at discovering and communicating about such differences. In addition, when the information *was* mentioned in the teams I studied, remote partners sometimes failed to note or remember it. This makes sense when we think of local information as a "hidden profile" (Stasser and Stewart 1992; Stasser et al. 1995; Stasser and Titus 1985, 1987). According to the principle of information sampling, uniquely held information is less likely than commonly held information to be mentioned in group discussions. If mentioned, it is less likely than commonly held information to be salient to group members. These problems are exacerbated by high information load (Stasser and Titus 1987) and use of text-based communication technologies (Hightower and Sayeed 1995, 1996; Hollingshead 1996). Thus, members of dispersed teams may have difficulty achieving mutual knowledge of important aspects of the situations and contexts in which partners function. In addition, the problem of unrecognized differences in context exacerbates other problems described below.

Unrecognized differences in context should be less of a problem for collocated teams than dispersed teams to the extent that collocated team members share context in common. Collocated teams also have more powerful ways of discovering differences, such as visual inspection and face-to-face communication. Future research should compare the mechanisms by which members of collocated and dispersed teams identify differences in situations, constraints, and assumptions. In addition, research should continue to examine how dispersed collaborators handle contextual information. It would be useful to identify conditions under which they are able to form mental maps of the situations of remote partners and update them as situations change. We may also wish to compare the ability of dispersed collaborators to detect differences in task-related information across locations relative to differences in context, including cultural context.

Communication across distance and via technology was shown in my study to be a particularly leaky process. Messages were addressed incorrectly, undelivered, or deliberately not sent to team members. People worked from different information far more often than they realized, and this caused serious problems in communication and relationships. Confusion and conflict was promulgated not just by different interpretations of the same information, but also by different interpretations of different information. I have shown that members often blamed each other for their frustrations.

This problem has implications for the development of trust in dispersed collaboration. Jarvenpaa and Leidner (1999) and Jarvenpaa et al. (1998) found that trust in dispersed teams was predicted most strongly during the early phases of team activity by perceptions of other members' "integrity," by which they mean "adherence to principles thought to make the trustee dependable and reliable" such

as demonstrated work ethic, fair dealings, and consistency" (Jarvenpaa et al. 1998, p. 31). The authors propose that trusting action and demonstrated reliability increase trust in dispersed teams. However, my work suggests that human and technical errors in information distribution may be common in dispersed collaboration, particularly during the early phases of activity. If these are interpreted as failures of personal reliability, they are likely to inhibit the development and maintenance of trust.

Failures of Interpretation

This study also identified three problems that disrupted shared interpretation of information in the dispersed teams: difficulty communicating and understanding the salience of information, differences in speed of access to information, and difficulty interpreting the meaning of silence. When problems of salience occurred, partners had the same information but attended to different parts of it and misunderstood each other as a result. While this problem is not unique to dispersed teams, it probably is exacerbated by the use of computer-mediated communication. The medium does not provide the paraverbal and verbal cues that people use in conversation to signal the importance of one piece of information relative to another. Furthermore, failure to communicate salience may be more costly in dispersed than collocated collaborations because of slow feedback channels and restricted backchannel feedback. Slow and effortful feedback limits detection and correction of misunderstandings. Kiesler and Sproull (1992, Sproull and Kiesler 1986) described how computer mediation restricts cues to the meaning of communication. I focus on the problem of signaling the importance of one piece of information relative to another and how this problem is manifested in dispersed collaboration and exacerbated by the other problems identified in this study.

The members of the dispersed teams I studied also had difficulty working together when the speed of feedback cycles differed among parts of the group. Some members were in frequent contact with rapid feedback cycles, while contact with others was limited and slower paced. These differences were caused by differences in access to communication technology and the distribution of members across distance and time. The teams were not synchronized in their access to information and ability to detect and correct misunderstandings, and so had a difficult time maintaining mutual knowledge.

This observation reveals a new side of the issue of rate in computer-mediated communication. Researchers have shown that the rate at which computer-mediated communication proceeds affects group productivity and the development of relationships (Straus 1997, Straus and McGrath 1994, Walther 1992, Walther and Burgoon 1992). My finding calls attention to the consequences of parts of a group communicating at different rates. Future research should explore whether uneven feedback cycles within a group have a different impact than uniform feedback cycles. Uneven feedback cycles across parts of a group could be more destructive than a uniformly slow pace because subgroups grow out of sync with, and isolated from, the group. This could result in their becoming scapegoats. Ironically, feedback cycles may be slower and more uneven among parts of a group under just those conditions for which rapid cycles are most needed: when the contexts of senders and receivers differ substantially. For example, feedback cycles may be unpredictable when part of a team is traveling constantly or located in an area with a weak communications infrastructure.

Finally, I observed that team members often misinterpreted the meaning of their remote partners' silence. Physical dispersion and dependence on communications technology add sources of uncertainty about the meaning of silence beyond those experienced by groups that meet face-to-face. Partners may fall silent because they find it difficult and time-consuming to convey sensitive issues in text, or because of technical failures. They may be silent because they agree, because they disagree, or because they are physically absent. In the mutual knowledge literature, Brennan (1998) describes how lack of feedback (i.e., silence) leads to failures of grounding in conversations with and through computers. Without feedback, one does not know whether a computer is working, has completed the task, has malfunctioned, or is waiting for additional inputs. I broaden her point by showing how physical dispersion presents additional sources of uncertainty as to the meaning of partners' silence.

Consequences for Attribution

The failures of information exchange and interpretation identified in this study have consequences for attribution processes. They illuminate two reasons why people are likely to make personal rather than situational attributions concerning remote partners. First, failure to share and remember information about remote situations and contexts, and uneven distribution of information, mean that remote partners often lack information to make situational attributions. According to the attribution literature, when people do not have situational information, they tend to make personal attributions, i.e., their explanations focus on the dispositions of individuals (Jones and Nisbett 1972, Nisbett et al. 1973).

The study also demonstrates the complexity of communicating and collaborating across distance and via

technology. Information about multiple locations must be gathered, integrated, and updated. Multiple possible explanations for unexpected behavior and silences must be weighed and investigated. Exchanges between subgroups must be reported to the whole. Feedback lags, which may be different for each location, must be taken into account. There is considerable evidence that when people work under heavy cognitive load, they become more likely to make personal rather than situational attributions (Gilbert and Hixon 1991, Gilbert and Osborne 1989, Gilbert et al. 1988).

Falling back on personal attributions because of a lack of information or information-processing limitations amounts to blaming individuals for problems that may have broader causes. This distracts partners from full diagnosis of problems and modification of practices to prevent reoccurrences. It also damages partners' opinions of each other. These points are consistent with the observations of Blakar (1984) and Hultberg et al. (1980) concerning the effects of personal versus situational and constructive versus nonconstructive attributions among the family members they studied.

Attribution processes among people who collaborate across distance and through the use of computer mediation merit additional attention. One of the contributions of this study is harnessing the power of the well developed literature concerning attribution to help understand the development of such relationships. Two information-based antecedents of attribution, situational information and cognitive load, are explored in this study. However, the attribution literature describes additional information-based antecedents of attribution, as well as a number of motivation-based antecedents. (See Kelley and Michaela 1980 for a summary.) This literature could help us understand how dispersed collaborators make sense of their complicated world.

Future research might also explore whether cognitive load is indeed higher in dispersed than collocated teams, and trace all its consequences. Two consequences are discussed in this paper: bias toward dispositional attribution and difficulty identifying uniquely held information. However, there could be other consequences over time, such as stress or burnout. In addition, the concept of cognitive load facilitates application of these findings to collocated teams: Members of collocated teams may be most likely to encounter the problems of mutual knowledge, and their consequences discussed here, when members are experiencing heavy cognitive load.

The data suggest that processes that began with failures of mutual knowledge and produced personal rather than situational attribution eventually led to the fracturing of some teams into in-groups and out-groups. There was a tendency to generalize attributions, particularly negative ones, to others at the same location. Team members' analysis papers describe remote *subgroups* as "lackadaisical," "aggressive," and having an "inferiority complex." This is consistent with the work of Lea and Spears (1991, 1992, 1993) who suggest that people using computer-mediated communication tend to categorize remote others on the basis of meager cues. In a dispersed team, one salient basis for social categorization is location, e.g., the California group or the Portuguese group or the client-site group. According to the literatures concerning group identity and in-group/out-group conflict, such tendencies are exacerbated by weak team integration (Karakowsky and Siegel 1995, McDonald 1995) and the need for a target for displaced hostility (Brewer 1986). Frustration in search of an outlet may build up in dispersed collaborations because of the elusive problems of information exchange and interpretation described in this study, and other structural and technical challenges. Once in-group/out-group dynamics had arisen in the teams I studied, subgroups tended to withhold information from each other. This erodes mutual knowledge to a greater degree, and worsens problems. It also creates differing impressions among parts of a group of the group's timing and pace, impacting motivation and coordination.

Future research should explore the role of social categorization processes in dispersed work groups, including generalization on the basis of location. Dynamics involving subgroups should be investigated because dispersed teams in practice typically include collocated subgroups (Goodman and Wilson 1998, Leonard et al. 1998, Maznevski and Chudoba 2000, Mazchrzak et al. 2000, Snow et al. 1996).

Association with Performance

There were no clear relationships between team performance and particular problems of information exchange and interpretation, or the general incidence of such problems. Failures of mutual knowledge were ubiquitous across the teams. Although it was possible to distinguish different team performance strategies, no one strategy was associated with high performance. However, performance strategies did seem to be associated with different *relational* outcomes in the teams. Future research should return to this issue with more sensitive task designs. In particular, designs should vary the distribution of task-related information across locations and the amount of interdependence required of team members. Using such tasks, relationships among team performance, performance strategies, and the failures of mutual knowledge identified in this study should be examined.

Amplifying and Moderating Forces

Feedback lags seem to amplify the problems of information exchange and interpretation identified by this study. Krauss and Bricker (1966) demonstrate that feedback lags disrupt the ability of senders and receivers to establish common referents, a building block of mutual knowledge. In addition, this study suggests that feedback lags contribute to the exaggeration of negative attributions concerning remote partners and make it more difficult for dispersed collaborators to diagnose their situation.

Without feedback, deprived collaborators are left to speculate why their expectations have not been fulfilled and when feedback will come. In the absence of situational information, they are likely to make negative attributions concerning the dispositions of their remote partners. These attributions can grow more negative as waiting continues. Deprived collaborators also sometimes amplify their demands, triggering an exaggerated response from their remote partners. The situation is like that of the person who ultimately turns the hot water up too high in the shower because of time lags between turns of the faucet and response. One participant said as much: "Some problems dragged on for days *while the suspicions of group members intensified*. In reality, the problem could have been as simple as someone not being able to get to the computer lab to check their messages." The speed of feedback cycles may constitute a critical constraint for geographically dispersed groups.

In addition, feedback lags and dispersed information make it extremely difficult for people to get an overview of the structure and functioning of a dispersed system of relationships. Actions and reactions are difficult to interpret when disrupted by lags in feedback. Blaming is a common response when individuals do not grasp the structure and dynamics of complex systems of which they are a part (Bowen 1985, Senge 1990). In future research, systems dynamics theory (Sterman 1989) might contribute to our understanding of the impacts of distributed information and feedback lags on dispersed collaboration and computer-mediated communication.

In both dispersed and collocated collaboration, problems establishing and maintaining mutual knowledge are most likely to occur when there is a great deal of uniquely held task-related and contextual information and limited communication channels. Exacerbating factors can be expected to include heavy cognitive load, a complex interdependent task, tight time limits, and a complex team design—particularly one involving strong subgroup identities, which may reinforce local perspectives.

For situations in which these factors are operating, practices that should moderate problems include methodically seeking out situational and uniquely held information, giving prompt feedback whenever possible, focusing on the overall structure and processes of the system of relationships rather than on individuals, reexamining group operating practices and norms, and extending the benefit of the doubt rather than engaging in the creation of out-groups. The overall effect of these practices is to direct attention to group-level diagnosis and learning.

Limitations

The mode of generalization appropriate to case study research is analytic generalization—generalization to theory rather than statistical generalization (Yin 1994). Therefore, it is important to articulate how the teams studied here may be typical and atypical of geographically dispersed work groups. Geographically dispersed work groups take many different forms in practice (Goodman and Wilson 1998, Leonard et al. 1998, Maznevski and Chudoba 2000, Mazchrzak et al. 2000, Snow et al. 1996). The teams I studied probably are atypical in the limitations they faced around means of communication. Travel, videoconferencing, and telephone conferencing were not an option for them, and they were limited by personal expense in their use of the telephone. Occasional face-to-face meetings and more telephone contact might moderate the processes observed; however, there is reason to think that basic tendencies might be the same. Additional modes of contact could contribute to uneven exchange of information among parts of a team if used extensively by dyads or subgroups.

Several team design factors should also be noted: group identity and time frame (discussed in Walther 1997), interdependence, and composition. The local university-based subgroups of the teams studied probably had a considerably stronger basis for identity than the teams as wholes. While dispersed teams in practice are often composed of people from multiple organizations and subgroups with strong identities, the weak basis for team identity should be taken into account. Likewise, the teams' seven-week time frame is not unusual in business practice; however, its relative shortness and the teams' low expectation of future interaction should also be noted. As discussed previously, the team's level of interdependence could be characterized as moderate. They carried out a project that required research, creativity, and a range of skills with outcomes of significance for the members. The teams were composed of adult professionals representing a range of ages with moderate international and technical experience. We must continue to examine the forms that dispersed collaboration takes in practice and

take these structural factors into account in our research designs and theories.

Another limitation of the study is the use of one primary judge of the meaning of the data. Case study and multiple case study methodologists struggle with the issue of reliability, given the unwieldy form of the data and the time required to review it, and this study faces the same challenge. This work followed the recommended practice of preserving a case study database and case study protocol (Yin 1994) so that another researcher could review the process. Following Eisenhardt (1989), the protocol included two reviews of all the cases that were aimed at testing reliability, one after initial formulation of constructs and the second after formulation of the model to refine or refute emerging conclusions.

There are also some safeguards peculiar to this study. There was a clear standard as to what constituted the foundation of the data because every effort was made to develop complete e-mail histories of each team. Communication and the flow of events were preserved in a record that did not reflect the choices of a researcher. While I was the final judge of the meaning of this communication and these events, I did compare my impressions with those recorded separately by a research assistant and by 26 participants in the teams. When these impressions diverged, I could and did minutely examine the e-mail records to try to understand the divergence.

In addition, I provided numerous detailed examples in this report so that some assessment of the evidence can be made in lieu of a full review of the records. Gersick (1988) observes that the advantage of a single patient judge of meaning is that analysis is done consistently, yielding understanding of a whole event. However, this will always be one side of a trade-off in a study such as this one.

A third limitation is that this study did not compare the dispersed teams with collocated teams, so we cannot determine the extent to which the problems and processes described also occur in collocated collaborations. This issue must be explored empirically. However, I have taken care to describe when we might see these dynamics in collocated collaborations, and to provide grounded reasons why failure to establish and maintain mutual knowledge should be more likely and more serious in dispersed collaborations.

Implications for Practice

While the conclusions offered here are tentative, designers and members of geographically dispersed teams may still be interested in the implications for practice. This work suggests that designers of dispersed teams should aggressively explore in advance potential differences in situations and incentives that will affect team members. Goals, incentives, and situations should be aligned whenever possible. When they cannot be, these differences should be brought to all team members' attention.

Ideally, all members of a dispersed collaboration should be sent the same information. Beyond the content of the information, this provides each member with an accurate picture of the *pace* of activity in the collaboration, including any differences in pace among subgroups. In practice, however, the information load could be overwhelming. If there is a risk of overload, leaders and members of dispersed teams should communicate information that establishes or makes adjustments to the parameters of collaboration such as (1) the availability of members (including identification of holidays), and constraints on availability such as competing responsibilities; (2) the objectives of the collaboration and solution contexts; (3) local requirements, customs, processes, and constraints that bear on member availability, objectives, or solutions; (4) means of communication and norms, including backup procedures; and (5) reports on the pace of activity overall and the pace in any subgroups.

Members of dispersed teams and people communicating via computer mediation should resist making assumptions about the situation and constraints of remote others. Instead, they should actively seek out such information. It is also important for individuals to monitor the tendency to leap to dispositional attributions about remote partners. Situational causes should be considered, even if information to support them is not immediately available. In addition, prompt feedback when possible helps everyone in a complex distributed system to correct inaccurate interpretations and attributions. Training in systems thinking may be useful for members of distributed work groups by helping them appreciate the structure, processes, and time lags of the system of which they are a part.

We have entered a new era of collaborative activity, one in which it is feasible for work groups to span time zones rather than yards or miles. There are many advantages to be gained through the use of such groups. However, their usefulness will be maximized if we understand characteristic dynamics sufficiently well, so that effective team designs can be developed and effective training can be offered. The literature on computer-mediated communication has led the way by exploring the nature of communication in such groups. However, it is argued here that not only the mode of communication but also the fact that these groups are complex distributed dynamic systems will affect processes, and outcomes. We know that people tend not to be sensitive to the structure, processes, and time lags of the systems of which they are

a part. However, successful collaboration in geographically dispersed work groups may hinge on members mastering these skills: grasping local realities and the big picture, establishing mutual knowledge, taking into account the dynamics of feedback processes, and understanding the processes linking distant but interdependent parts.

Acknowledgments

The author thanks *Organization Science* Senior Editor Wanda Orlikowski for guiding the review and refinement of this paper with great wisdom. The author also appreciates the helpful comments provided by Katherine Chudoba, Gerardine DeSanctis, Pamela Hinds, George Huber, Richard Klimoski, David Kravitz, Fred Niederman, Roger Volkema, Joseph Walther, and three anonymous reviewers. An early version of this paper appeared in the 1997 Academy of Management *Best Paper Proceedings* and benefited from the comments of three anonymous reviewers.

The author is grateful to Beata Lobert Jones, who organized the Virtual Learning Environment project and shared records with her; her research assistant Colleen Lambert, who intelligently organized and summarized thousands of e-mails; and her graduate business students, who embraced the opportunity to explore a frontier of collaboration and share their adventures.

References

Arthur, M., D. Rousseau. 1996. A new career lexicon for the 21st century. *Acad. Management Executive* **10**(4) 28–39.

Armstrong, D., P. Cole. 1995. Managing distances and differences in geographically distributed work groups. S. Jackson, M. Ruderman, eds. *Diversity in Work Teams*. American Psychological Association, Washington, DC 187–216.

Blakar, R. 1973. An experimental method for inquiring into communication. *European J. Social Psych.* **3** 415–425.

——. 1984. *Communication: A Social Perspective on Clinical Issues*. Columbia, New York.

——. 1985. Towards a theory of communication in terms of preconditions: A conceptual framework and some empirical explorations. H. Giles, R. St. Clair, eds. *Recent Advances in Language, Communication, and Social Psychology*. Lawrence Erlbaum, London, U.K. 10–40.

Bowen, M. 1985. *Family Therapy in Clinical Practice*. Jason Aronson, Northvale, NJ.

Boudreau, M., K. Loch, D. Robey, D. Straud. 1998. Going global: Using information technology to advance the competitiveness of the virtual transnational organization. *Acad. Management Executive* **12**(4) 120–128.

Brennan, S. 1998. The grounding problem in conversations with and through computers. S. Fussell, R. Kreuz, eds. *Social and Cognitive Approaches to Interpersonal Communication*. Lawrence Erlbaum, Hillsdale, NJ 210–225.

Brewer, M. 1986. The role of ethnocentrism in intergroup conflict. W. Worchel, W. Austin, eds. *Psychology of Intergroup Relations*. Nelson, Chicago, IL 88–102.

Clark, H. 1996. *Using Language*. Cambridge University Press, New York.

——, T. Carlson. 1982. Speech acts and hearer's beliefs. N. Smith, ed. *Mutual Knowledge*. Academic Press, New York 1–36.

——, C. Marshall. 1981. Definite reference and mutual knowledge. A. Joshi, I. Sag, B. Webber, eds. *Elements of Discourse Understanding*: 10–63. Cambridge University Press, New York.

Coulam, R., R. Smith. 1985. Introduction. R. Coulam, R. Smith, eds. *Advances in Information Processing in Organizations.* JAI Press, Stamford, CT Vol. 2, 1–19.

Dennis, A. 1996. Information exchange and use in small group decision making. *Small Group Res*. **27**(4) 532–549.

——, S. Pootheri, V. Natarajan. 1996. *http://tcbworks.cba.uga.edu/.*

DeSanctis, G., M. Poole. 1994. Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organ. Sci.* **5**(2) 121–147.

——, ——. 1997. Transitions in teamwork in new organizational forms. *Advances in Group Processes*. JAI Press, Stamford, CT Vol. 14, 157–176.

Eisenhardt, K. 1989. Building theories from case study research. *Acad. Management Rev.* **14**(4) 532–550.

Elsbach, K., R. Sutton. 1992. Acquiring organizational legitimacy through illegitimate actions: A marriage of institutional and impression management theories. *Acad. Management J.* **35** 699–738.

Fussell, S., R. Krauss. 1992. Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *J. Personality and Social Psych.* **62**(3) 378–391.

Gersick, C. 1988. Time and transition in work teams: Toward a new model of group development. *Acad. Management J.* **31** 9–41.

Gilbert, D., J. Hixon. 1991. The trouble with thinking: Activation and application of stereotypic beliefs. *J. Personality and Social Psych.* **60**(4) 509–517.

——, R. Osborne. 1989. Thinking backward: Some curable and incurable consequences of cognitive busyness. *J. Personality and Social Psych.* **57**(6) 940–949.

——, B. Pelham, D. Krull. 1988. On cognitive busyness: When person perceivers meet persons perceived. *J. Personality and Social Psych.* **54**(5) 733–740.

Goodman, P., J. Wilson. (in press). The changing nature and role of teams. T. Kochan, ed. *Negotiations and Change: From Workplace to Society.*

Handy, C. 1995. Trust and the virtual organization. *Harvard Bus. Rev.*, **73**(May-June) 40–50.

Heider, F. 1958. *The Psychology of Interpersonal Relations*. Wiley, New York.

Hightower, R., L. Sayeed. 1995. The impact of computer-mediated communication systems on biased group discussion. *Computers in Organizational Behavior* **11**(1) 33–44.

——, ——. 1996. Effects of communication mode and prediscussion information distribution characteristics on information exchange in groups. *Inform. Systems Res.* **7**(4) 451–465.

Hollingshead, A. 1996. Information suppression and status persistence in group decision making: The effects of communication media. *Human Commun. Res.* **23**(2) 193–219.

Hultberg, M., S. Alve, R. Blakar. 1980. Patterns of attribution of communication difficulties in couples with a 'borderline,' a 'schizophrenic,' or a 'normal' offspring. R. Blakar, ed. *Studies of Familial Communication and Psychopathology: A Social-Developmental Approach to Deviant Behavior*. Columbia University Press, New York 66–104.

Jackson, J. 1965. Structural characteristics of norms. D. Steiner, M.

Fishbein, eds. *Current Studies in Social Psychology*. Holt, Rinehart and Winston, New York.

Jarvenpaa, S., K. Knoll, D. Leidner. 1998. Is anybody out there? Antecedents of trust in global virtual teams. *J. Management Inform. Systems* **14**(4) 29–64.

——, D. Leidner. 1999. Communication and trust in global virtual teams. *Organ. Sci.* **10**(6) 791–815.

Jones, E., R. Nisbett. 1972. The actor and the observer: Divergent perceptions of the causes of behavior. E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, B. Weiner, eds. *Attribution: Perceiving the Causes of Behavior*. General Learning Press, Morristown, NJ 79–94.

Karakowsky, L., J. Siegel. 1995. The effect of demographic diversity on causal attributions of work group success and failure: A framework for research. M. Martinko, ed. *Attribution Theory: An Organizational Perspective*. St. Lucie Press, Delray Beach, FL **77** 289–313.

Kelly, H., J. Michela. 1980. Attribution theory and research. *Annual Rev. Psych.* **31** 457–501.

Kemske, F. 1998. HR 2008. *Workforce* (January) 47–60.

Kiesler, S., J. Siegel, T. McGuire. 1984. Social psychological aspects of computer-mediated communication. *Amer. Psych.* **39**(10) 1123–1134.

——, L. Sproull. 1992. Group decision making and communication technology. *Organ. Behavior and Human Decision Processes* **52** 96–123.

——, D. Zubrow, A. Moses, V. Geller. 1985. Affect in computer-mediated communication: An experiment in synchronous terminal-to-terminal discussion. *Human Comput. Interaction* **1** 77–104.

Krauss, R., P. Bricker. 1966. Effects of transmission delay and access delay on the efficiency of verbal communication. *J. Acoustical Society* **41** 286–292.

——, S. Fussell. 1990. Mutual knowledge and communicative effectiveness. J. Galegher, R. Kraut, C. Egido, Eds. *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*. Lawrence Erlbaum, Hillsdale, NJ 111–146.

——, S. Lewis, L. Swezey. 1982. Listener responsiveness and the coordination of conversation. *J. Personality and Soc. Psych.* **43**(4) 718–731.

Lea, M., R. Spears. 1991. Computer-mediated communication, deindividuation and group decision-making. *Internat. J. Man-Machine Stud.* **34** 283–301.

——, ——. 1992. Paralanguage and social perception in computer-mediated communication. *J. Organ. Comput.* **2** 321–341.

——, ——. 1993. Love at first byte? Building personal relationships over computer networks. J. Wood, S. Duck, eds. *Under-studied Relationships*. Sage, Thousand Oaks, CA 197–233.

Lebie, L., J. Rhoades, J. McGrath. 1996. Interaction process in computer-mediated and face-to-face groups. *Computer Supported Cooperative Work* **4** 127–152.

Leonard, D., P. Brands, A. Edmondson, J. Fenwick. 1998. Virtual teams: Using communications technology to manage geographically dispersed development groups. S. Bradley, R. Nolan, eds. *Sense and Respond: Capturing Value in the Network Era*. Harvard Business School Press, Cambridge, MA 285–298.

Lipnack, J., J. Stamps. 1997. *Virtual Teams: Reaching across Space, Time and Organizations with Technology*. Wiley, New York.

Majchrzak, A., R. Rice, N. King, A. Malhotra, S. Ba. 2000. Computer-mediated inter-organizational knowledge-sharing: Insights from a virtual team innovating using a collaborative tool. *Inform. Resources Management J.* **13**(Jan–Feb) 44–53.

Maznevski, M., K. Chudoba. 2000. Bridging space over time: Global virtual team dynamics and effectiveness. *Organ. Sci.* **11**(5) 473–492.

McDonald, D. 1995. Fixing blame in n-person attributions: A social identity model for attributional processes in newly formed cross-functional groups. M. Martinko, ed. *Attribution Theory: An Organizational Perspective*. St. Lucie Press, Delray Beach, FL 273–288.

McGrath, J., A. Hollingshead. 1994. *Groups Interacting with Technology*. Sage, Thousand Oaks, CA.

McGuire, T., S. Kiesler, J. Siegel. 1987. Group and computer-mediated discussion effects in risk decision making. *J. Personality and Social Psych.* **52** 917–930.

McLeod, P. 1996. New communication technologies for group decision-making: Toward an integrative framework. R. Hirokawa, M. Poole, eds. *Communication and Group Decision Making*, 2nd ed. Sage, Thousand Oaks, CA 426–461.

Miles, M., A. Huberman. 1994. *Qualitative Data Analysis*, 2nd ed. Sage, Thousand Oaks, CA.

Nisbett, R., C. Caputo, P. Legant, J. Marecek. 1973. Behavior as seen by the actor and as seen by the observer. *J. Personality and Social Psych.* **27**(2) 154–164.

O'Hara-Devereaux, M., R. Johansen. 1994. *Globalwork: Bridging Distance, Culture and Time*. Jossey Bass, San Francisco, CA.

Redding, W. 1972. *Communication Within the Organization*. Industrial Communication Council, New York.

Rommetveit, R. 1968. *Words, Meanings and Messages*. Academic Press, New York.

Senge, P. 1990. *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday/Currency, New York.

Siegel, J., V. Dubrovsky, S. Kiesler, T. McGuire. 1986. Group process in computer-mediated communication. *Organ. Behavior and Human Decision Processes* **37** 157–187.

Snow, C., S. Snell, S. Davison, D. Hambrick. 1996. Use transnational teams to globalize your company. *Organ. Dynam.* **24**(4) 50–76.

Sproull, L., S. Kiesler. 1986. Reducing social context cues: Electronic mail in organizational communication. *Management Sci.* **32**(11) 1492–1512.

Stasser, G., D. Stewart. 1992. The discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *J. Personality and Social Psych.* **63** 426–434.

——, ——, G. Wittenbaum. 1995. Expert roles and information exchange during discussion: The importance of knowing who knows what. *J. Experiment. Social Psych.* **31** 244–265.

——, W. Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *J. Personality and Social Psych.* **48** 1467–1478.

——, ——. 1987. Effects of information load and percentages of shared information on the dissemination of unshared information during group discussion. *J. Personality and Social Psych.* **53** 81–93.

Sterman, J. 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Sci.* **35**(3) 321–339.

Straus, S. 1997. Technology, group process, and group outcomes: Testing the connections in computer-mediated and face-to-face groups. *Human-Computer Interaction*. **12** 227–266.

——, J. McGrath. 1994. Does the medium matter? The interaction of task type and technology on group performance and member reactions. *J. Appl. Psych.* **79**(2) 87–97.

Tannen, D. 1994. *Talking from 9 to 5*. William Morrow, New York.

Townsend, A., S. DeMarie, A. Hendrickson. 1998. Virtual teams: Technology and the workplace of the future. *Acad. Management Executive* **12**(3) 17–29.

Walther, J. 1992. Interpersonal effects in computer-mediated interaction: A relational perspective. *Commun. Res.* **19**(1) 52–90.

Accepted by Wanda Orlikowski.

——. 1997. Group and interpersonal effects in international computer-mediated collaboration. *Human Commun. Res.* **23**(3) 342–369.

——, J. Burgoon. 1992. Relational communication in computer-mediated interaction. *Human Commun. Res.* **19**(1) 50–88.

Warkentin, M., L. Sayeed, R. Hightower. 1997. Virtual teams versus face-to-face teams: An exploratory study of a web-based conference system. *Decision Sci.* **28**(4) 975–996.

Weisband, S. 1992. Group discussion and first advocacy effects in computer-mediated and face-to-face decision making groups. *Organ. Behavior and Human Decision Processes* **53** 352–380.

Yin, R. 1994. *Case Study Research* 2nd ed. Sage, Thousand Oaks, CA.

Yngve, V. 1970. On getting a word in edgewise. *Papers from the Sixth Regional Meeting, Chicago Linguistics Circle.* Chicago Linguistics Circle, Chicago, IL.

THE CHRONICLE
of Higher Education

Previous
← The Redesigned Evernote iPhone App

Next
Checking in with DHAnswers →

12 Basic Principles of Project Management

March 3, 2011, 11:00 am
By Brian Croxall

Last week I attended a two-day training session on the fundamentals of project management. I had been looking forward to the opportunity as a chance to help me be more effective at my current alt-ac job, where I work in a team environment on several very cool projects (a THATCamp, a digital scholarship commons, some library games). Working in a team has been a relatively new experience for me. After all, graduate school—at least in the humanities—doesn't often teach that skill set. For this reason, I thought it could be useful to share some of my training, in the same way that I did when I learned how to speak to the news media at the most recent MLA. Here, then, are twelve basic things that I learned about working with and leading a team:

- **Projects are temporary.** You can't consider a project to be that thing that you're going to do every day for the rest of your career. Instead, it's something that creates a particular product or service, and it has a clear end point. You might compare it to the creation of a syllabus or teaching a course. It has a finite beginning and end.
- **Decide whether or not the project *should* happen.** Not every project should be begun started. When you are beginning work on something, you want to determine if it's a good use of your resources, what problem the project is trying to solve, and whether or not the project is the best way to fix it. I think that this is particularly hard in universities because we tend to originate our own projects rather than having them brought to us (more on this below). It can be very difficult to admit that what you've been wanting to do isn't worth the time, money, and/or effort. But we have to be willing to call a spade a spade.
- **Consider risks.** If you're going to tackle the project after all, you should analyze your potential risks. What events might derail the project? What are the likelihood of them happening? Working together to brainstorm what these things will help you plan for the events that seem most likely or severe. For example, your entire digital humanities project could be derailed by a server failure. For this reason, you should probably be obsessive about backup.
- **Cost, time, and quality are co-dependent.** In other words, once a project has been started, you can't change its timeline without directly affecting its cost or quality. You can't expect to get something done faster without either paying more *or* sacrificing some quality. This is why getting a plumber to your house at 10pm costs more than waiting for the next day. The same thing applies to any project you undertake with a team.
- **Know what's out of bounds.** Determining what your project *will not* do is just as important as determining what it will since that can help prevent mission creep in the future.
- **Develop a project plan with clear activities.** At the beginning of a project, the size or scope of the task might seem overwhelming. (Ask anyone who's ever tried writing a book!) However, breaking things into small and manageable chunks can help you eat that proverbial elephant. When working on a team, you should have these chunks start with a verb so your team members clearly know what you're asking them to do.
- **When making assignments, consider people's interest as much as their skills and experience.** Just because someone has a ton of experience designing websites doesn't mean that it's the only thing that he wants to do. Letting people choose how they want to be involved in the project allows them to develop personally as well as helping the project.
- **Let the person taking an assignment set the due date.** It can be hard to manage a project and not come off as an ogre at times. But

one way to shed the Shrek this is to let those who are getting the assignments decide when they can complete them. Their estimates won't always be 100% accurate, but they will not have the excuse of it being a deadline that is imposed on them. Moreover, getting a team member's input helps them feel more connected to the project.

- **There are lots of project management tools; just use what works for you.** This is pretty self-explanatory and is the subject of this great thread at DHAnswers (see previous ProfHacker coverage here, here, and here). If post-its and emails work for your team, then go for it. If you need to use Basecamp and can get others to do so, that's fine too. In the end, it's not the tool so much as the relationships that count for being successful at a project. As our instructor said, "A fool with a tool is still a fool."
- **There can only be one.** In a progressive, neo-liberal environment (read, university), we tend to want to let everyone get involved in decision making. In the end, however, you can't really share project management duties. One person needs to be chief. But it also doesn't have to be you all the time.
- **Set meeting ground rules.** It's true that bad meetings are your fault—especially if you're the project manager. Getting your team to agree collectively to how meetings will work will help things run smoothly in the future. And be sure to only hold meetings when they are necessary.
- **Celebrate success.** Since projects are finite, they *will* have an end...no matter how far off that might seem at present. When you get to this point, make sure that you figure out some way to celebrate the accomplishment. A celebration doesn't have to cost a lot of money—or any—but recognizing others' contributions and the completion of the goal is important.

That's certainly not everything that I learned in 2 days, but I think it's the key points that I'll be taking forward with me. A lot of it is common sense, but I'll be the first to admit that PhDs aren't always gifted with common sense (otherwise they might not begin the degree in the first place!).

It's worth mentioning that not everything we covered seemed germane to working in a university. The most important of these is that the concepts of project management (insofar as they were introduced to me) appear linked to doing something *for someone else*. What I mean by this is that project management is designed to help you meet the needs of a client who comes and asks you for something and who will be judging whether or not you have succeeded. In many ways, then, it's a form of consulting within a top-down structure.

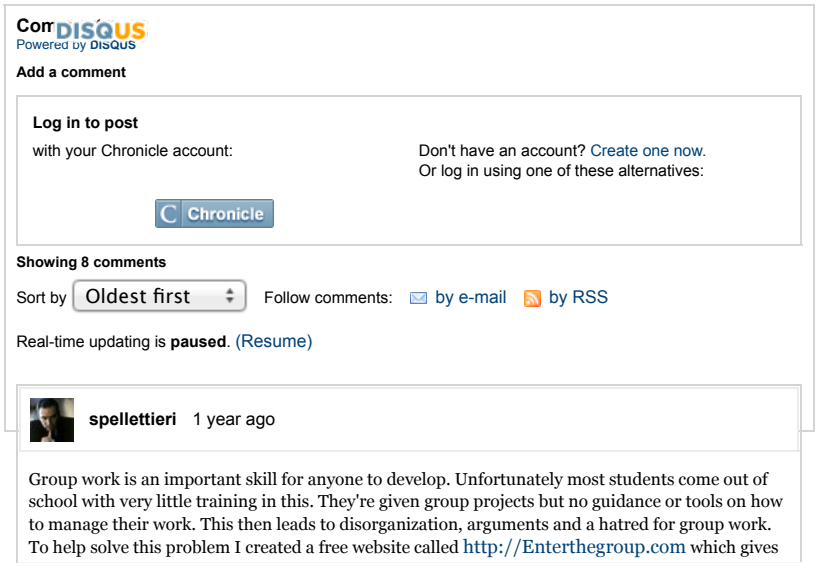
But in universities, I think, ideas and projects tend to percolate *up* from individuals. It's normally a faculty member, graduate student, librarian, or someone else who has an idea, gets working on it, and eventually pulls in a team to help. In this case, the initiator of the project is both the client and the project manager. The distinction between these two types of clients isn't just academic. (Ahem.) It can be hard, for instance, to create a budget for a project when you are working to put together funding from, say, different campus departments, as I did when I began a lecture series. You create a budget, of course, to show people; but it's not the same thing as being told by a client that her company can spend $25,000 on project X. So some of the steps and considerations that a project manager working for IBM might make simply don't translate to the university.

This disjunction between "regular" businesses and universities certainly does not make project management useless. In fact, I've got some great ideas. But it is also true that we must adapt these principles so we can better manage team members in our own specific contexts.

To get more particularly at that context, then, **what's the single most important lesson you've learned about working in a team or on a project?** Please let us know in the comments.

Lead image: Meeting nieuwe leden / Voka Kamer van Koophandel Limburg / CC BY 2.0

This entry was posted in Productivity, Profession and tagged jargon, meetings, project, project management, projects, teams. Bookmark the permalink.



students tools and advice on how to work more effectively in groups.

Like

DelaneyKirk 1 year ago

Thanks for the article. I'm teaching a module on project management in my grad class this semester so I'll be sharing this info.

1 person liked this. Like

chapmandw 1 year ago

Very nice piece... particularly, "There can only be one." Someone has to be in charge, whether it's 2 people or 200.

Like

Micah Vandegrift 1 year ago

I'm still brand, brand new at project management, but I'm learning a ton right off the bat. All the points you listed above have come up at one time or another in the 6 months I've been working on this project. The one thing I have learned is that eventually decisions have to be made. I'm working on a digitization project, and spent a good 2 months researching DAMS systems, and eventually just had to pick one and run with it. Also, depending on the type of project, the "There can only be one" point can get really complicated. Thanks for these tips though. Project management is one skill that I wish I had been taught rather than having to learn/teach myself.

Like

ccsucio 1 year ago

A useful summary of key points. I suggest two additions, PM evaluation and reflective assessment. First, I find it useful to gather feedback about the just-completed PM process using a "what worked" and "what didn't" approach. Second, and along the same line is an assessment of the project implementation 6-12 months later. This review doesn't have to be extensive but follow-up helps us learn more about the process and helps reinforce the value of the project. The only caveat here is that, like all PM, you have to be selective (e.g. Large projects; projects over a certain amount of $$$) about such reviews as they can be time-consuming. I have found that in higher ed only a handful of institutions have full-time PMs even in IT. Project Management is a skill that needs to be developed in all managers.

1 person liked this. Like

englishlibrarian 1 year ago

I would add: make sure that you have buy-in from your entire team. A single member who is lukewarm about the project can sabotage the entire enterprise!

Like

DelaneyKirk 1 year ago

Check out Timothy Johnson's book "Race Through the Forest." It's written as a business fable and is an interesting way to teach project management.

Like

pvaughn93 9 months ago

loved it

1 person liked this. Like

Copyright 2012. All rights reserved.

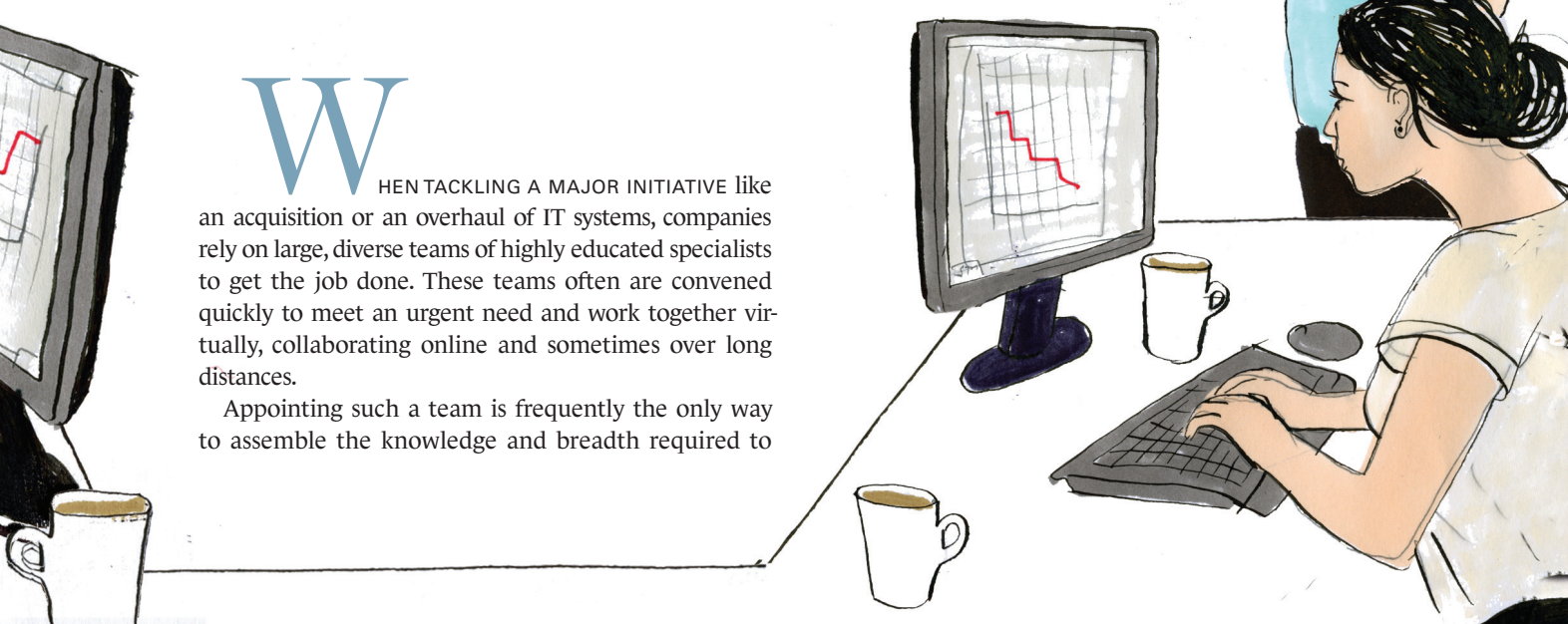
The Chronicle of Higher Education 1255 Twenty-Third St, N.W. Washington, D.C. 20037

Clare Mallison

8 Ways to Build Collaborative Teams

Even the largest and most complex teams can work together effectively if the right conditions are in place.

by Lynda Gratton and Tamara J. Erickson



WHEN TACKLING A MAJOR INITIATIVE like an acquisition or an overhaul of IT systems, companies rely on large, diverse teams of highly educated specialists to get the job done. These teams often are convened quickly to meet an urgent need and work together virtually, collaborating online and sometimes over long distances.

Appointing such a team is frequently the only way to assemble the knowledge and breadth required to

pull off many of the complex tasks businesses face today. When the BBC covers the World Cup or the Olympics, for instance, it gathers a large team of researchers, writers, producers, cameramen, and technicians, many of whom have not met before the project. These specialists work together under the high pressure of a "no retake" environment, with just one chance to record the action. Similarly, when the central IT team at Marriott sets out to develop sophisticated systems to enhance guest experiences, it has to collaborate closely with independent hotel owners, customer-experience experts, global brand managers, and regional heads, each with his or her own agenda and needs.

Our recent research into team behavior at 15 multinational companies, however, reveals an interesting paradox: Although teams that are large, virtual, diverse, and composed of highly educated specialists are increasingly crucial with challenging projects, those same four characteristics make it hard for teams to get anything done. To put it another way, the qualities required for success are the same qualities that undermine success. Members of complex teams are less likely – *absent other influences* – to share knowledge freely, to learn from one another, to shift workloads flexibly to break up unexpected bottlenecks, to help one another complete jobs and meet deadlines, and to share resources – in other words, to collaborate. They are less likely to say that they "sink or swim" together, want one another to succeed, or view their goals as compatible.

Consider the issue of size. Teams have grown considerably over the past ten years. New technologies help companies extend participation on a project to an ever greater number of people, allowing firms to tap into a wide body of knowledge and expertise. A decade or so ago, the common view was that true teams rarely had more than 20 members. Today, according to our research, many complex tasks involve teams of 100 or more. However, as the size of a team increases beyond 20 members, the tendency to collaborate naturally decreases, we have found. Under the right conditions, large teams can achieve high levels of cooperation, but creating those conditions requires thoughtful, and sometimes significant, investments in the capacity for collaboration across the organization.

Working together virtually has a similar impact on teams. The majority of those we studied had members spread among multiple locations – in several cases, in as many as 13 sites around the globe. But as teams became more virtual, we saw, cooperation also declined, unless the company had taken measures to establish a collaborative culture.

As for diversity, the challenging tasks facing businesses today almost always require the input and expertise of people with disparate views and backgrounds to create cross-fertilization that sparks insight and innovation. But diversity also creates problems. Our research shows that team members collaborate more easily and naturally if they perceive themselves as being alike. The differences that inhibit collaboration include not only nationality but also age, educational level, and even tenure. Greater diversity also often means that team members are working with people that they know only superficially or have never met before – colleagues drawn from other divisions of the company, perhaps, or even from outside it. We have found that the higher the proportion of strangers on the team and the greater the diversity of background and experience, the less likely the team members are to share knowledge or exhibit other collaborative behaviors.

We found that the greater the proportion of experts a team had, the more likely it was to disintegrate into nonproductive conflict or stalemate.

In the same way, the higher the educational level of the team members is, the more challenging collaboration appears to be for them. We found that the greater the proportion of experts a team had, the more likely it was to disintegrate into nonproductive conflict or stalemate.

So how can executives strengthen an organization's ability to perform complex collaborative tasks – to maximize the effectiveness of large, diverse teams, while minimizing the disadvantages posed by their structure and composition?

To answer that question we looked carefully at 55 large teams and identified those that demonstrated high levels

Lynda Gratton (lgratton@london.edu) is a professor of management practice at London Business School and a senior fellow at the Advanced Institute of Management. She is the author of *Hot Spots: Why Some Teams, Workplaces, and Organizations Buzz with Energy – and Others Don't* (Berrett-Koehler, 2007). **Tamara J. Erickson** (tjerickson@concoursgroup.com) is the president of the Concours Institute, the research and education arm of BSG Alliance. She is based in Boston and is a coauthor of several articles for HBR, including the McKinsey Award winner "It's Time to Retire Retirement" (March 2004).

Collaboration Conundrums

Four traits that are crucial to teams – but also undermine them

Large Size

Whereas a decade ago, teams rarely had more than 20 members, our findings show that their size has increased significantly, no doubt because of new technologies. Large teams are often formed to ensure the involvement of a wide stakeholder group, the coordination of a diverse set of activities, and the harnessing of multiple skills. As a consequence, many inevitably involve 100 people or more. However, our research shows that as the size of the team increases beyond 20 members, the level of natural cooperation among members of the team decreases.

Virtual Participation

Today most complex collaborative teams have members who are working at a distance from one another. Again, the logic is that the assigned tasks require the insights and knowledge of people from many locations. Team members may be working in offices in the same city or strung across the world. Only 40% of the teams in our sample had members all in one place. Our research shows that as teams become more virtual, collaboration declines.

Diversity

Often the challenging tasks facing today's businesses require the rapid assembly of people from multiple backgrounds and perspectives, many of whom have rarely, if ever, met. Their diverse knowledge and views can spark insight and innovation. However, our research shows that the higher the proportion of people who don't know anyone else on the team and the greater the diversity, the less likely the team members are to share knowledge.

High Education Levels

Complex collaborative teams often generate huge value by drawing on a variety of deeply specialized skills and knowledge to devise new solutions. Again, however, our research shows that the greater the proportion of highly educated specialists on a team, the more likely the team is to disintegrate into unproductive conflicts.

of collaborative behavior despite their complexity. Put differently, they succeeded both because of and despite their composition. Using a range of statistical analyses, we considered how more than 100 factors, such as the design of the task and the company culture, might contribute to collaboration, manifested, for example, in a willingness to share knowledge and workloads. Out of the 100-plus factors, we were able to isolate eight practices that correlated with success – that is, that appeared to help teams overcome substantially the difficulties that were posed by size, long-distance communication, diversity, and specialization. We then interviewed the teams that were very strong in these practices, to find out how they did it. In this article we'll walk through the practices. They fall into four general categories – executive support, HR practices, the strength of the team leader, and the structure of the team itself.

Executive Support

At the most basic level, a team's success or failure at collaborating reflects the philosophy of top executives in the organization. Teams do well when executives invest in supporting social relationships, demonstrate collaborative behavior themselves, and create what we call a "gift culture" – one in which employees experience interactions with leaders and colleagues as something valuable and generously offered, a gift.

Investing in signature relationship practices. When we looked at complex collaborative teams that were performing in a productive and innovative manner, we found that in every case the company's top executives had invested significantly in building and maintaining social relationships throughout the organization. However, the way they did that varied widely. The most collaborative companies had what we call "signature" practices – practices that were memorable, difficult for others to replicate, and particularly well suited to their own business environment.

For example, when Royal Bank of Scotland's CEO, Fred Goodwin, invested £350 million to open a new headquarters building outside Edinburgh in 2005, one of his goals was to foster productive collaboration among employees. Built around an indoor atrium, the new structure allows more than 3,000 people from the firm to rub shoulders daily.

The headquarters is designed to improve communication, increase the exchange of ideas, and create a sense of community among employees. Many of the offices have an open layout and look over the atrium – a vast transparent space. The campus is set up like a small town, with retail shops, restaurants, jogging tracks and cycling trails, spaces for picnics and barbecues – even a leisure club complete with swimming pool, gym, dance studios, tennis courts, and football pitches. The idea is that with a private "Main Street" running through the headquarters, employees will remain on the campus throughout the day – and be out of their offices mingling with colleagues for at least a portion of it.

To ensure that non-headquarters staff members feel they are a part of the action, Goodwin also commissioned an adjoining business school, where employees from other locations meet and learn. The visitors are encouraged to spend time on the headquarters campus and at forums designed to give employees opportunities to build relationships.

Indeed, the RBS teams we studied had very strong social relationships, a solid basis for collaborative activity that allowed them to accomplish tasks quickly. Take the Group Business Improvement (GBI) teams, which work on 30-, 60-, or 90-day projects ranging from back-office fixes to IT updates and are made up of people from across RBS's many businesses, including insurance, retail banking, and private

banking in Europe and the United States. When RBS bought NatWest and migrated the new acquisition's technology platform to RBS's, the speed and success of the GBI teams confounded many market analysts.

BP has made another sort of signature investment. Because its employees are located all over the world, with relatively few at headquarters, the company aims to build social networks by moving employees across functions, businesses, and countries as part of their career development. When BP integrates an acquisition (it has grown by buying numerous smaller oil companies), the leadership development committee deliberately rotates employees from the acquired firm through positions across the corporation. Though the easier and cheaper call would be to leave the executives in their own units – where, after all, they know the business – BP instead trains them to take on new roles. As a consequence any senior team today is likely to be made up of people from multiple heritages. Changing roles frequently – it would not be uncommon for a senior leader at BP to have worked in four businesses and three geographic locations over the past decade – forces executives to become very good at meeting new people and building relationships with them.

Modeling collaborative behavior. In companies with many thousands of employees, relatively few have the opportunity to observe the behavior of the senior team on a day-to-day basis. Nonetheless, we found that the perceived behavior of senior executives plays a significant role in determining how cooperative teams are prepared to be.

Executives at Standard Chartered Bank are exceptionally good role models when it comes to cooperation, a strength that many attribute to the firm's global trading heritage. The Chartered Bank received its remit from Queen Victoria in 1853. The bank's traditional business was in cotton from Bombay (now Mumbai), indigo and tea from Calcutta, rice from Burma, sugar from Java, tobacco from Sumatra, hemp from Manila, and silk from Yokohama. The Standard Bank was founded in the Cape Province of South Africa in 1863 and was prominent in financing the development of the diamond fields and later gold mines. Standard Chartered was formed in 1969 through a merger of the two banks, and today the firm has 57 operating groups in 57 countries, with no home market.

It's widely accepted at Standard Chartered that members of the general management committee will frequently serve as substitutes for one another. The executives all know and understand the entire business and can fill in for each other easily on almost any task, whether it's leading a regional celebration, representing the company at a key external event, or kicking off an internal dialogue with employees.

While the behavior of the executive team is crucial to supporting a culture of collaboration, the challenge is to make executives' behavior visible. At Standard Chartered the senior team travels extensively; the norm is to travel even

Eight Factors That Lead to Success

1 **Investing in signature relationship practices.** Executives can encourage collaborative behavior by making highly visible investments – in facilities with open floor plans to foster communication, for example – that demonstrate their commitment to collaboration.

2 **Modeling collaborative behavior.** At companies where the senior executives demonstrate highly collaborative behavior themselves, teams collaborate well.

3 **Creating a "gift culture."** Mentoring and coaching – especially on an informal basis – help people build the networks they need to work across corporate boundaries.

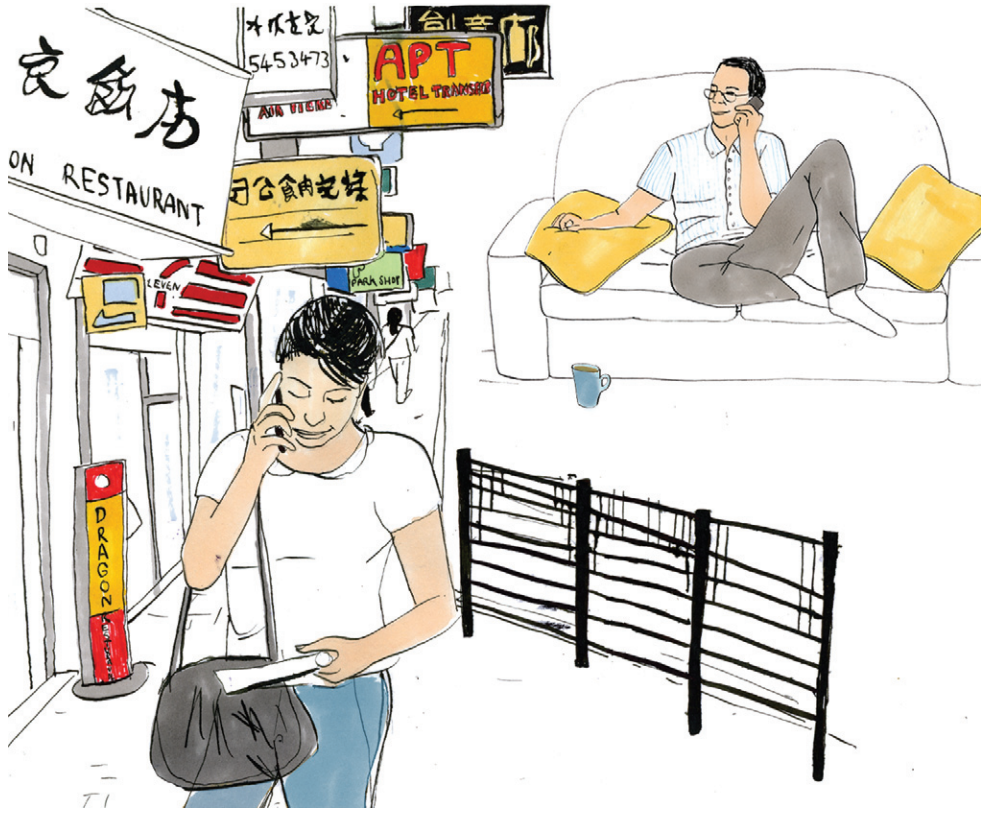
4 **Ensuring the requisite skills.** Human resources departments that teach employees how to build relationships, communicate well, and resolve conflicts creatively can have a major impact on team collaboration.

5 **Supporting a strong sense of community.** When people feel a sense of community, they are more comfortable reaching out to others and more likely to share knowledge.

6 **Assigning team leaders that are both task- and relationship-oriented.** The debate has traditionally focused on whether a task or a relationship orientation creates better leadership, but in fact both are key to successfully leading a team. Typically, leaning more heavily on a task orientation at the outset of a project and shifting toward a relationship orientation once the work is in full swing works best.

7 **Building on heritage relationships.** When too many team members are strangers, people may be reluctant to share knowledge. The best practice is to put at least a few people who know one another on the team.

8 **Understanding role clarity and task ambiguity.** Cooperation increases when the roles of individual team members are sharply defined yet the team is given latitude on how to achieve the task.



for relatively brief meetings. This investment in face-to-face interaction creates many opportunities for people across the company to see the top executives in action. Internal communication is frequent and open, and, maybe most telling, every site around the world is filled with photos of groups of executives – country and functional leaders – working together.

The senior team's collaborative nature trickles down throughout the organization. Employees quickly learn that the best way to get things done is through informal networks. For example, when a major program was recently launched to introduce a new customer-facing technology, the team responsible had an almost uncanny ability to understand who the key stakeholders at each branch bank were and how best to approach them. The team members' first-name acquaintance with people across the company brought a sense of dynamism to their interactions.

Creating a "gift culture." A third important role for executives is to ensure that mentoring and coaching become embedded in their own routine behavior – and throughout the company. We looked at both formal mentoring processes, with clear roles and responsibilities, and less formal processes, where mentoring was integrated into everyday activities. It turned out that while both types were important, the latter was more likely to increase collaborative behavior. Daily coaching helps establish a cooperative "gift culture" in place of a more transactional "tit-for-tat culture."

At Nokia informal mentoring begins as soon as someone steps into a new job. Typically, within a few days, the employee's manager will sit down and list all the people in the organization, no matter in what location, it would be useful for the employee to meet. This is a deeply ingrained cultural norm, which probably originated when Nokia was a smaller and simpler organization. The manager sits with the newcomer, just as her manager sat with her when she joined, and reviews what topics the newcomer should discuss with each person on the list and why establishing a relationship with him or her is important. It is then standard for the newcomer to actively set up meetings with the people on the list, even when it means traveling to other locations. The gift of time – in the form of hours spent on coaching and building networks – is seen as crucial to the collaborative culture at Nokia.

Focused HR Practices

So what about human resources? Is collaboration solely in the hands of the executive team? In our study we looked at the impact of a wide variety of HR practices, including selection, performance management, promotion, rewards, and training, as well as formally sponsored coaching and mentoring programs.

We found some surprises: for example, that the type of reward system – whether based on team or individual achievement, or tied explicitly to collaborative behavior or not – had no discernible effect on complex teams' productivity and innovation. Although most formal HR programs appeared to have limited impact, we found that two practices did improve team performance: training in skills related to collaborative behavior, and support for informal community building. Where collaboration was strong, the HR team had typically made a significant investment in one or both of those practices – often in ways that uniquely represented the company's culture and business strategy.

Ensuring the requisite skills. Many of the factors that support collaboration relate to what we call the "container" of collaboration – the underlying culture and habits of the company or team. However, we found that some teams had a collaborative culture but were not skilled in the practice of collaboration itself. They were encouraged to cooperate,

they wanted to cooperate, but they didn't know how to work together very well in teams.

Our study showed that a number of skills were crucial: appreciating others, being able to engage in purposeful conversations, productively and creatively resolving conflicts, and program management. By training employees in those areas, a company's human resources or corporate learning department can make an important difference in team performance.

In the research, PricewaterhouseCoopers emerged as having one of the strongest capabilities in productive collaboration. With responsibility for developing 140,000 employees in nearly 150 countries, PwC's training includes modules that address teamwork, emotional intelligence, networking, holding difficult conversations, coaching, corporate social responsibility, and communicating the firm's strategy and shared values. PwC also teaches employees how to influence others effectively and build healthy partnerships.

A number of other successful teams in our sample came from organizations that had a commitment to teaching employees relationship skills. Lehman Brothers' flagship program for its client-facing staff, for instance, is its training in selling and relationship management. The program is not about sales techniques but, rather, focuses on how Lehman values its clients and makes sure that every client has access to all the resources the firm has to offer. It is essentially a course on strategies for building collaborative partnerships with customers, emphasizing the importance of trust-based personal relationships.

Supporting a sense of community. While a communal spirit can develop spontaneously, we discovered that HR can also play a critical role in cultivating it, by sponsoring group events and activities such as women's networks, cooking weekends, and tennis coaching, or creating policies and practices that encourage them.

At ABN Amro we studied effective change-management teams within the company's enterprise services function. These informal groups were responsible for projects associated with the implementation of new technology throughout the bank; one team, for instance, was charged with expanding online banking services. To succeed, the teams needed the involvement and expertise of different parts of the organization.

The ABN Amro teams rated the company's support for informal communities very positively. The firm makes the technology needed for long-distance collaboration readily available to groups of individuals with shared interests – for instance, in specific technologies or markets – who hold frequent web conferences and communicate actively online. The company also encourages employees that travel to a new location to arrange meetings with as many people as possible. As projects are completed, working groups disband but employees maintain networks of connections. These practices serve to build a strong community over time – one that sets the stage for success with future projects.

The most productive, innovative teams were led by people who were *both* task- and relationship-oriented. What's more, these leaders changed their style during the project.

Committed investment in informal networks is also a central plank of the HR strategy at Marriott. Despite its size and global reach, Marriott remains a family business, and the chairman, Bill Marriott, makes a point of communicating that idea regularly to employees. He still tells stories of counting sticky nickels at night as a child – proceeds from the root-beer stand founded in downtown Washington, DC, by his mother and father.

Many of the firm's HR investments reinforce a friendly, family-like culture. Almost every communication reflects an element of staff appreciation. A range of "pop-up" events – spontaneous activities – create a sense of fun and community. For example, the cafeteria might roll back to the 1950s, hold a twist dance contest, and in doing so, recognize the anniversary of the company's first hotel opening. Bill Marriott's birthday might be celebrated with parties throughout the company, serving as an occasion to emphasize the firm's culture and values. The chairman recently began his own blog, which is popular with employees, in which he discusses everything from Marriott's efforts to become greener, to his favorite family vacation spots – themes intended to reinforce the idea that the company is a community.

The Right Team Leaders

In the groups that had high levels of collaborative behavior, the team leaders clearly made a significant difference. The question in our minds was how they actually achieved this. The answer, we saw, lay in their flexibility as managers.

Assigning leaders who are both task- and relationship-oriented. There has been much debate among both academ-

ics and senior managers about the most appropriate style for leading teams. Some people have suggested that relationship-oriented leadership is most appropriate in complex teams, since people are more likely to share knowledge in an environment of trust and goodwill. Others have argued that a task orientation – the ability to make objectives clear, to create a shared awareness of the dimensions of the task, and to provide monitoring and feedback – is most important.

In the 55 teams we studied, we found that the truth lay somewhere in between. The most productive, innovative teams were typically led by people who were *both* task- and relationship-oriented. What's more, these leaders changed their style during the project. Specifically, at the early stages they exhibited task-oriented leadership: They made the goal clear, engaged in debates about commitments, and clarified the responsibilities of individual team members. However, at a certain point in the development of the project they switched to a relationship orientation. This shift often took place once team members had nailed down the goals and their accountabilities and when the initial tensions around sharing knowledge had begun to emerge. An emphasis throughout a project on one style at the expense of the other inevitably hindered the long-term performance of the team, we found.

Producing ambidextrous team leaders – those with both relationship and task skills – is a core goal of team-leadership development at Marriott. The company's performance-review process emphasizes growth in both kinds of skills. As evidence of their relationship skills, managers are asked to describe their peer network and cite examples of specific ways that network helped them succeed. They also must provide examples of how they've used relationship building to get things done. The development plans that follow these conversations explicitly map out how the managers can improve specific elements of their social relationships and networks. Such a plan might include, for instance, having lunch regularly with people from a particular community of interest.

To improve their task leadership, many people in the teams at Marriott participated in project-management certification programs, taking refresher courses to maintain their skills over time. Evidence of both kinds of capabilities becomes a significant criterion on which people are selected for key leadership roles at the company.

Team Formation and Structure

The final set of lessons for developing and managing complex teams has to do with the makeup and structure of the teams themselves.

Building on heritage relationships. Given how important trust is to successful collaboration, forming teams that capitalize on preexisting, or "heritage," relationships, increases the chances of a project's success. Our research shows that new teams, particularly those with a high proportion of members who were strangers at the time of formation, find it more difficult to collaborate than those with established relationships.

Newly formed teams are forced to invest significant time and effort in building trusting relationships. However, when some team members already know and trust one another, they can become nodes, which over time evolve into networks. Looking closely at our data, we discovered that when 20% to 40% of the team members were already well connected to one another, the team had strong collaboration right from the start.

It helps, of course, if the company leadership has taken other measures to cultivate networks that cross boundaries. The orientation process at Nokia ensures that a large number of people on any team know one another, increasing the odds that even in a company of more than 100,000 people, someone on a companywide team knows someone else and can make introductions.

Nokia has also developed an organizational architecture designed to make good use of heritage relationships. When it needs to transfer skills across business functions or units, Nokia moves entire small teams intact instead of reshuffling individual people into new positions. If, for example, the company needs to bring together a group of market and technology experts to address a new customer need, the group formed would be composed of small pods of

How Complex Is the Collaborative Task?

Not all highly collaborative tasks are complex. In assembling and managing a team, consider the project you need to assign and whether the following statements apply:

- ❑ The task is unlikely to be accomplished successfully using only the skills within the team.
- ❑ The task must be addressed by a new group formed specifically for this purpose.
- ❑ The task requires collective input from highly specialized individuals.
- ❑ The task requires collective input and agreement from more than 20 people.
- ❑ The members of the team working on the task are in more than two locations.
- ❑ The success of the task is highly dependent on understanding preferences or needs of individuals outside the group.
- ❑ The outcome of the task will be influenced by events that are highly uncertain and difficult to predict.
- ❑ The task must be completed under extreme time pressure.

If more than two of these statements are true, the task requires complex collaboration.

colleagues from each area. This ensures that key heritage relationships continue to strengthen over time, even as the organization redirects its resources to meet market needs. Because the entire company has one common platform for logistics, HR, finance, and other transactions, teams can switch in and out of businesses and geographies without learning new systems.

One important caveat about heritage relationships: If not skillfully managed, too many of them can actually disrupt collaboration. When a significant number of people within the team know one another, they tend to form strong subgroups – whether by function, geography, or anything else they have in common. When that happens, the probability of conflict among the subgroups, which we call fault lines, increases.

Understanding role clarity and task ambiguity. Which is more important to promoting collaboration: a clearly defined approach toward achieving the goal, or clearly specified roles for individual team members? The common assumption is that carefully spelling out the approach is essential, but leaving the roles of individuals within the team vague will encourage people to share ideas and contribute in multiple dimensions.

Our research shows that the opposite is true: Collaboration improves when the roles of individual team members are clearly defined and well understood – when individuals feel that they can do a significant portion of their work independently. Without such clarity, team members are likely to waste too much energy negotiating roles or protecting turf, rather than focus on the task. In addition, team members are more likely to want to collaborate if the path to achieving the team's goal is left somewhat ambiguous. If a team perceives the task as one that requires creativity, where the approach is not yet well known or predefined, its members are more likely to invest time and energy in collaboration.

At the BBC we studied the teams responsible for the radio and television broadcasts of the 2006 Proms (a two-month-long musical celebration), the team that televised the 2006 World Cup, and a team responsible for daytime television news. These teams were large – 133 people worked on the Proms, 66 on the World Cup, and 72 on the news – and included members with a wide range of skills and from many disciplines. One would imagine, therefore, that there was a strong possibility of confusion among team members.

To the contrary, we found that the BBC's teams scored among the highest in our sample with regard to the clarity with which members viewed their own roles and the roles of others. Every team was composed of specialists who had deep expertise in their

The Research

Our work is based on a major research initiative conducted jointly by the Concours Institute (a member of BSG Alliance) and the Cooperative Research Project of London Business School, with funding from the Advanced Institute for Management and 15 corporate sponsors. The initiative was created as a way to explore the practicalities of collaborative work in contemporary organizations.

We sent surveys to 2,420 people, including members of 55 teams. A total of 1,543 people replied, a response rate of 64%. Separate surveys were administered to group members, to group leaders, to the executives who evaluated teams, and to HR leaders at the companies involved. The tasks performed by the teams included new-product development, process reengineering, and identifying new solutions to business problems. The companies involved included four telecommunication companies, seven financial services or consulting firms, two media companies, a hospitality firm, and one oil company. The size of the teams ranged from four to 183 people, with an average of 44.

Our objective was to study the levers that executives could pull to improve team performance and innovation in collaborative tasks. We examined scores of possible factors, including the following:

The general culture of the company. We designed a wide range of survey questions to measure the extent to which the firm had a cooperative culture and to uncover employees' attitudes toward knowledge sharing.

Human resources practices and processes. We studied the way staffing took place and the process by which people were promoted. We examined the extent and type of training, how reward systems were configured, and the extent to which mentoring and coaching took place.

Socialization and network-building practices. We looked at how often people within the team participated in informal socialization, and the type of interaction that was most common. We also asked numerous questions about the extent to which team members were active in informal communities.

The design of the task. We asked team members and team leaders about the task itself. Our interest here was in how they perceived the purpose of the task, how complex it was, the extent to which the task required members of the team to be interdependent, and the extent to which the task required them to engage in boundary-spanning activities with people outside the team.

The leadership of the team. We studied the perceptions team members had of their leaders' style and how the leaders described their own style. In particular, we were interested in the extent to which the leaders practiced relationship-oriented and task-oriented skills and set cooperative or competitive goals.

The behavior of the senior executives. We asked team members and team leaders about their perceptions of the senior executives of their business unit. We focused in particular on whether team members described them as cooperative or competitive.

In total we considered more than 100 factors. Using a range of statistical analyses, we were able to identify eight that correlated with the successful performance of teams handling complex collaborative tasks. (See the sidebar "Eight Factors That Lead to Success.")

given function, and each person had a clearly defined role. There was little overlap between the responsibilities of the sound technician and the camera operator, and so on. Yet the tasks the BBC teams tackle are, by their very nature, uncertain, particularly when they involve breaking news. The trick the BBC has pulled off has been to clarify team members' individual roles with so much precision that it keeps friction to a minimum.

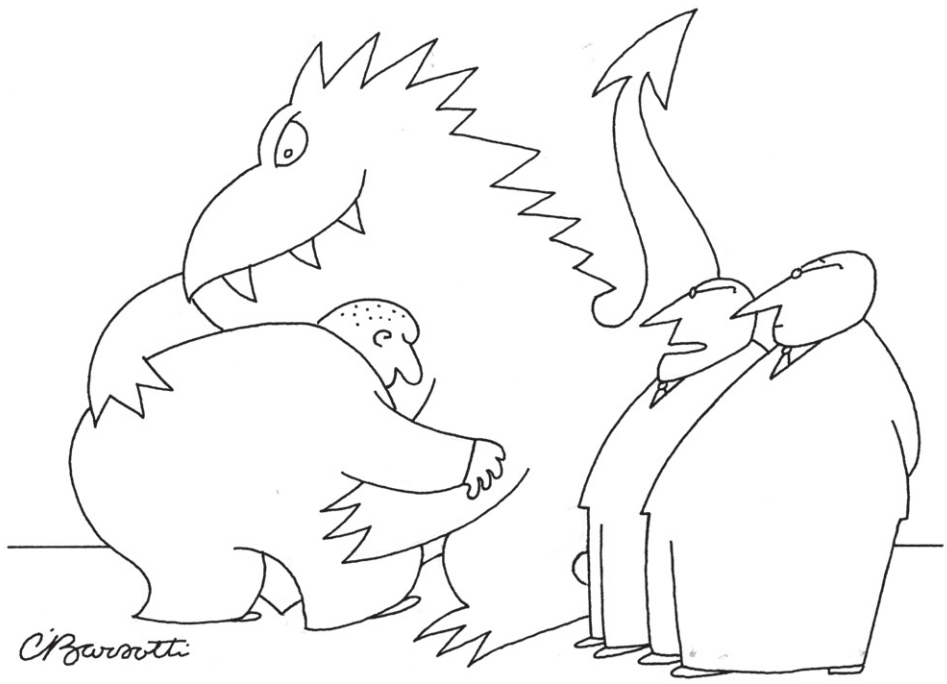
The successful teams we studied at Reuters worked out of far-flung locations, and often the team members didn't speak a common language. (The primary languages were Russian, Chinese, Thai, and English.) These teams, largely composed of software programmers, were responsible for the rapid development of highly complex technical software and network products. Many of the programmers sat at their desks for 12 hours straight developing code, speaking with no one. Ironically, these teams judged cooperative behavior to be high among their members. That may be because each individual was given autonomy over one discrete piece of the project. The rapid pace and demanding project timelines encouraged individual members to work independently to get the job done, but each person's work had to be shaped with an eye toward the overall team goal.

•••

Strengthening your organization's capacity for collaboration requires a combination of long-term investments – in building relationships and trust, in developing a culture in which senior leaders are role models of cooperation – and smart near-term decisions about the ways teams are formed, roles are defined, and challenges and tasks are articulated. Practices and structures that may have worked well with simple teams of people who were all in one location and knew one another are likely to lead to failure when teams grow more complex.

Most of the factors that impede collaboration today would have impeded collaboration at any time in history. Yesterday's teams, however, didn't require the same amount of members, diversity, long-distance cooperation, or expertise that teams now need to solve global business challenges. So the models for teams need to be realigned with the demands of the current business environment. Through careful attention to the factors we've described in this article, companies can assemble the breadth of expertise needed to solve complex business problems – without inducing the destructive behaviors that can accompany it.

Reprint R0711F
To order, see page 155.



Charles Barsotti

"One of my private demons – I hired Louis to wrestle it for me."

Harvard Business Review Notice of Use Restrictions, May 2009

Harvard Business Review and Harvard Business Publishing Newsletter content on EBSCOhost is licensed for the private individual use of authorized EBSCOhost users. It is not intended for use as assigned course material in academic institutions nor as corporate learning or training materials in businesses. Academic licensees may not use this content in electronic reserves, electronic course packs, persistent linking from syllabi or by any other means of incorporating the content into course resources. Business licensees may not host this content on learning management systems or use persistent linking or other means to incorporate the content into learning management systems. Harvard Business Publishing will be pleased to grant permission to make this content available through such means. For rates and permission, contact permissions@harvardbusiness.org.

1993

It won't surprise anyone to find an article on teams by Jon Katzenbach and Douglas Smith figuring into an issue devoted to high performance. While Peter Drucker may have been the first to point out that a team-based organization can be highly effective, Katzenbach and Smith's work made it possible for companies to implement the idea.

In this groundbreaking 1993 article, the authors say that if managers want to make better decisions about teams, they must be clear about what a team is. They define a team as "a small number of people with complementary skills who are committed to a common purpose, set of performance goals, and approach for which they hold themselves mutually accountable." That definition lays down the discipline that teams must share to be effective.

Katzenbach and Smith discuss the four elements – common commitment and purpose, performance goals, complementary skills, and mutual accountability – that make teams function. They also classify teams into three varieties – teams that recommend things, teams that make or do things, and teams that run things – and describe how each type faces different challenges.

The Discipline of Teams

by Jon R. Katzenbach and Douglas K. Smith

What makes the difference between a team that performs and one that doesn't?

Early in the 1980s, Bill Greenwood and a small band of rebel railroaders took on most of the top management of Burlington Northern and created a multibillion-dollar business in "piggybacking" rail services despite widespread resistance, even resentment, within the company. The Medical Products Group at Hewlett-Packard owes most of its leading performance to the remarkable efforts of Dean Morton, Lew Platt, Ben Holmes, Dick Alberding, and a handful of their colleagues who revitalized a health care business that most others had written off. At Knight Ridder, Jim Batten's "customer obsession" vision took root at the *Tallahassee Democrat* when 14 frontline enthusiasts turned a charter to eliminate errors into a mission of major change and took the entire paper along with them.

Such are the stories and the work of teams – real teams that perform, not amorphous groups that we call teams because we think that the label is motivating and energizing. The difference between teams that perform and other groups that don't is a subject to which most of us pay far too little attention. Part of the problem is that "team" is a word and concept so familiar to everyone. (See the exhibit "Not All Groups Are Teams: How to Tell the Difference.")

Or at least that's what we thought when we set out to do research for our book *The Wisdom of Teams* (HarperBusiness, 1993). We wanted to discover what differentiates various levels of

team performance, where and how teams work best, and what top management can do to enhance their effectiveness. We talked with hundreds of people on more than 50 different teams in 30 companies and beyond, from Motorola and Hewlett-Packard to Operation Desert Storm and the Girl Scouts.

We found that there is a basic discipline that makes teams work. We also found that teams and good performance are inseparable: You cannot have one without the other. But people use the word "team" so loosely that it gets in the way of learning and applying the discipline that leads to good performance. For managers to make better decisions about whether, when, or how to encourage and use teams, it is important to be more precise about what a team is and what it isn't.

Most executives advocate teamwork. And they should. Teamwork represents a set of values that encourage listening and responding constructively to views expressed by others, giving others the benefit of the doubt, providing support, and recognizing the interests and achievements of others. Such values help teams perform, and they also promote individual performance as well as the performance of an entire organization. But teamwork values by themselves are not exclusive to teams, nor are they enough to ensure team performance. (See the sidebar "Building Team Performance.")

Nor is a team just any group working together. Committees, councils, and task forces are not necessarily teams. Groups do not become teams simply because that is what someone calls them. The entire workforce of any large and complex organization is *never* a team, but think about how often that platitude is offered up.

To understand how teams deliver extra performance, we must distinguish between teams and other forms of working groups. That distinction turns on performance results. A working group's performance is a function of what its members do as individuals. A team's performance includes both individual results and what we call "collective work products." A collective work product is what two or more members must work on together, such as interviews, surveys, or experiments. Whatever it is, a collective work product reflects the joint, real contribution of team members.

Working groups are both prevalent and effective in large organizations where individual accountability is most important. The best working groups come together to share information, perspectives, and insights; to make decisions that help each person do his or her job better; and to reinforce individual performance standards. But the focus is always on individual goals and accountabilities. Working-group members don't take responsibility for results other than their own. Nor do they try to develop incremental performance contributions requiring the combined work of two or more members.

Teams differ fundamentally from working groups because they require both individual and mutual accountability. Teams rely on more than group discussion, debate, and decision, on more than sharing information and best-practice performance standards. Teams produce discrete work products through the joint contributions of their members. This is what makes possible performance levels greater than the sum of all the individual bests of team members. Simply stated, a team is more than the sum of its parts.

The first step in developing a disciplined approach to team management is to think about teams as discrete units of performance and not just as positive sets of values. Having observed and worked with scores of teams in action, both successes and failures, we offer the following. Think of it as a working defi-

Not All Groups Are Teams: How to Tell the Difference

Working Group	Team
> Strong, clearly focused leader	> Shared leadership roles
> Individual accountability	> Individual and mutual accountability
> The group's purpose is the same as the broader organizational mission	> Specific team purpose that the team itself delivers
> Individual work products	> Collective work products
> Runs efficient meetings	> Encourages open-ended discussion and active problem-solving meetings
> Measures its effectiveness indirectly by its influence on others (such as financial performance of the business)	> Measures performance directly by assessing collective work products
> Discusses, decides, and delegates	> Discusses, decides, and does real work together

Jon R. Katzenbach *is a founder and senior partner of Katzenbach Partners, a strategic and organizational consulting firm, and a former director of McKinsey & Company. His most recent book is* Why Pride Matters More Than Money: The Power of the World's Greatest Motivational Force *(Crown Business, 2003).* ***Douglas K. Smith*** *is an organizational consultant and a former partner at McKinsey & Company. His most recent book is* On Value and Values: Thinking Differently About We in an Age of Me *(Financial Times Prentice Hall, 2004).*

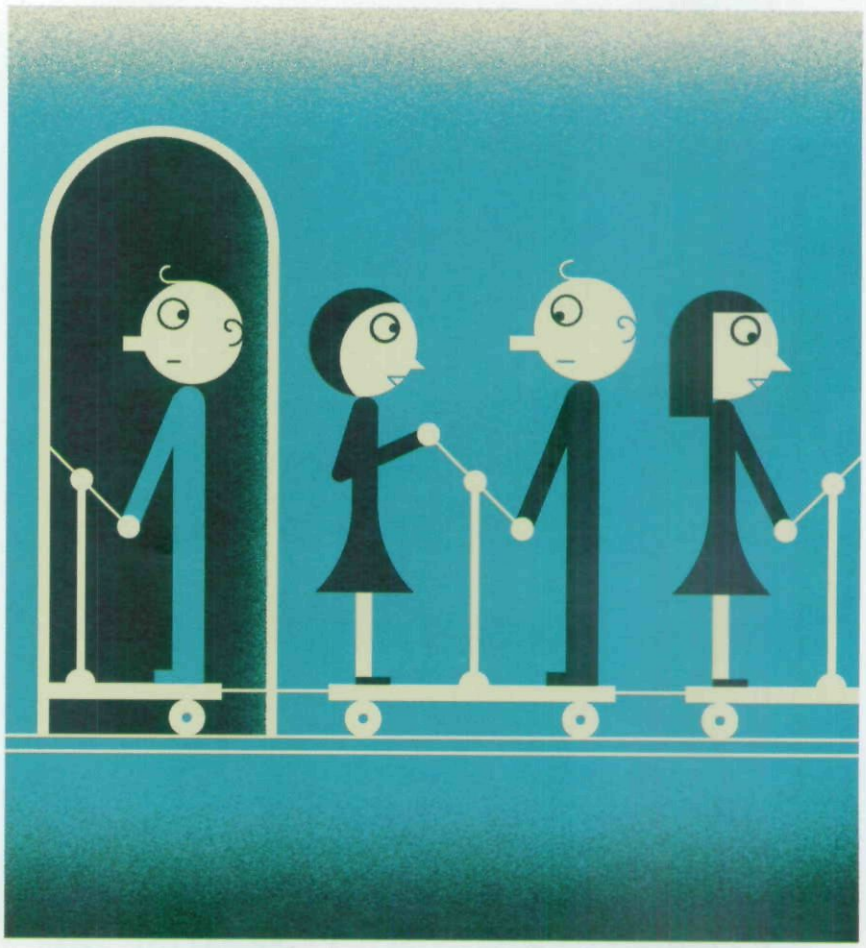
nition or, better still, an essential discipline that real teams share: *A team is a small number of people with complementary skills who are committed to a common purpose, set of performance goals, and approach for which they hold themselves mutually accountable.*

The essence of a team is common commitment. Without it, groups perform as individuals; with it, they become a powerful unit of collective performance. This kind of commitment requires a purpose in which team members can believe. Whether the purpose is to "transform the contributions of suppliers into the satisfaction of customers," to "make our company one we can be proud of again," or to "prove that all children can learn," credible team purposes have an element related to winning, being first, revolutionizing, or being on the cutting edge.

Teams develop direction, momentum, and commitment by working to shape a meaningful purpose. Building ownership and commitment to team purpose, however, is not incompatible with taking initial direction from outside the team. The often-asserted assumption that a team cannot "own" its purpose unless management leaves it alone actually confuses more potential teams than it helps. In fact, it is the exceptional case – for example, entrepreneurial situations–when a team creates a purpose entirely on its own.

Most successful teams shape their purposes in response to a demand or opportunity put in their path, usually by higher management. This helps teams get started by broadly framing the company's performance expectation. Management is responsible for clarifying the charter, rationale, and performance challenge for the team, but management must also leave enough flexibility for the team to develop commitment around its own spin on that purpose, set of specific goals, timing, and approach.

The best teams invest a tremendous amount of time and effort exploring, shaping, and agreeing on a purpose that belongs to them both collectively and individually. This "purposing" activity continues throughout the life of the team. By contrast, failed teams rarely develop a common purpose. For whatever reason – an insufficient focus on performance, lack of effort, poor leadership – they do not coalesce around a challenging aspiration.



GREG MABLY

People use the word "team" so loosely that it gets in the way of learning and applying the discipline that leads to good performance.

The best teams also translate their common purpose into specific performance goals, such as reducing the reject rate from suppliers by 50% or increasing the math scores of graduates from 40% to 95%. Indeed, if a team fails to establish specific performance goals or if those goals do not relate directly to the team's overall purpose, team members become confused, pull apart, and revert to mediocre performance. By contrast, when purposes and goals build on one another and are combined with team commitment, they become a powerful engine of performance.

Transforming broad directives into specific and measurable performance goals is the surest first step for a team trying to shape a purpose meaningful to its members. Specific goals, such as getting a new product to market in less than half the normal time, responding to all customers within 24 hours, or achieving a zero-defect rate while simultaneously cutting costs by 40%, all provide firm footholds for teams. There are several reasons:

- Specific team-performance goals help define a set of work products that are different both from an organization-wide

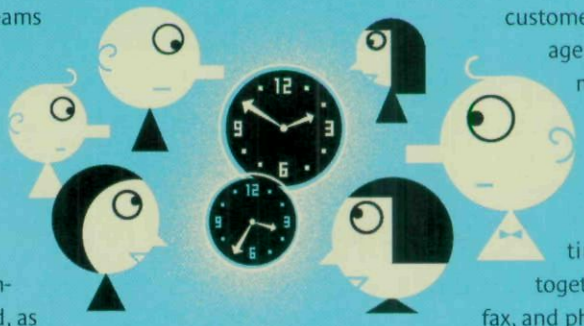
Building Team Performance

Although there is no guaranteed how-to recipe for building team performance, we observed a number of approaches shared by many successful teams.

Establish urgency, demanding performance standards, and direction. All team members need to believe the team has urgent and worthwhile purposes, and they want to know what the expectations are. Indeed, the more urgent and meaningful the rationale, the more likely it is that the team will live up to its performance potential, as was the case for a customer-service team that was told that further growth for the entire company would be impossible without major improvements in that area. Teams work best in a compelling context. That is why companies with strong performance ethics usually form teams readily.

Select members for skill and skill potential, not personality. No team succeeds without all the skills needed to meet its purpose and performance goals. Yet most teams figure out the skills they will need after they are formed. The wise manager will choose people for their existing skills and their potential to improve existing skills and learn new ones.

Pay particular attention to first meetings and actions. Initial impressions always mean a great deal. When potential teams first gather, everyone monitors the signals given by others to confirm, suspend, or dispel assumptions and concerns. They pay particular attention to those in authority: the team leader and any executives who set up, oversee, or otherwise influence the team. And, as always, what such leaders do is more important than what they say. If a senior executive leaves the team kickoff to take a phone call ten minutes after the session has begun and he never returns, people get the message.



Set some clear rules of behavior. All effective teams develop rules of conduct at the outset to help them achieve their purpose and performance goals. The most critical initial rules pertain to attendance (for example, "no interruptions to take phone calls"), discussion ("no sacred cows"), confidentiality ("the only things to leave this room are what we agree on"), analytic approach ("facts are friendly"), end-product orientation ("everyone gets assignments and does them"), constructive confrontation ("no finger pointing"), and, often the most important, contributions ("everyone does real work").

Set and seize upon a few immediate performance-oriented tasks and goals. Most effective teams trace their advancement to key performance-oriented events. Such events can be set in motion by immediately establishing a few challenging goals that can be reached early on. There is no such thing as a real team without performance results, so the sooner such results occur, the sooner the team congeals.

Challenge the group regularly with fresh facts and information. New information causes a team to redefine and enrich its understanding of the performance challenge, thereby helping the team shape a common purpose, set clearer goals, and improve its common approach. A plant quality improvement team knew the cost of poor quality was high, but it wasn't until they researched the different types of defects and put a price tag on each one that they knew where to go next. Conversely, teams err when they assume that all the information needed exists in the collective experience and knowledge of their members.

Spend lots of time together. Common sense tells us that team members must spend a lot of time together, scheduled and unscheduled, especially in the beginning. Indeed, creative insights as well as personal bonding require impromptu and casual interactions just as much as analyzing spreadsheets and interviewing customers. Busy executives and managers too often intentionally minimize the time they spend together. The successful teams we've observed all gave themselves the time to learn to be a team. This time need not always be spent together physically; electronic, fax, and phone time can also count as time spent together.

Exploit the power of positive feedback, recognition, and reward. Positive reinforcement works as well in a team context as elsewhere. Giving out "gold stars" helps shape new behaviors critical to team performance. If people in the group, for example, are alert to a shy person's initial efforts to speak up and contribute, they can give the honest positive reinforcement that encourages continued contributions. There are many ways to recognize and reward team performance beyond direct compensation, from having a senior executive speak directly to the team about the urgency of its mission to using awards to recognize contributions. Ultimately, however, the satisfaction shared by a team in its own performance becomes the most cherished reward.

mission and from individual job objectives. As a result, such work products require the collective effort of team members to make something specific happen that, in and of itself, adds real value to results. By contrast, simply gathering from time to time to make decisions will not sustain team performance.

• The specificity of performance objectives facilitates clear communication and constructive conflict within the team. When a plant-level team, for example, sets a goal of reducing average machine changeover time to two hours, the clarity of the goal forces the team to concentrate on what it would take either to achieve or to reconsider the goal. When such goals are clear, discussions can focus on how to pursue them or whether to change them; when goals are ambiguous or nonexistent, such discussions are much less productive.

• The attainability of specific goals helps teams maintain their focus on getting results. A product-development team at Eli Lilly's Peripheral Systems Division set definite yardsticks for the market introduction of an ultrasonic probe to help doctors locate deep veins and arteries. The probe had to have an audible signal through a specified depth of tissue, be capable of being manufactured at a rate of 100 per day, and have a unit cost less than a preestablished amount. Because the team could measure its progress against each of these specific objectives, the team knew throughout the development process where it stood. Either it had achieved its goals or not.

• As Outward Bound and other team-building programs illustrate, specific objectives have a leveling effect conducive to team behavior. When a small group of people challenge themselves to get over a wall or to reduce cycle time by 50%, their respective titles, perks, and other stripes fade into the background. The teams that succeed evaluate what and how each individual can best contribute to the team's goal and, more important, do so in terms of the performance objective itself rather than a person's status or personality.

• Specific goals allow a team to achieve small wins as it pursues its broader purpose. These small wins are invaluable to building commitment and overcoming the inevitable obstacles that get in the way of a long-term purpose. For example, the Knight Ridder team mentioned at the outset turned a narrow goal to eliminate errors into a compelling customer service purpose.

• Performance goals are compelling. They are symbols of accomplishment that motivate and energize. They challenge the people on a team to commit themselves, as a team, to make a difference. Drama, urgency, and a healthy fear of failure combine to drive teams that have their collective eye on an attainable, but challenging, goal. Nobody but the team can make it happen. It's their challenge.

The combination of purpose and specific goals is essential to performance. Each depends on the other to remain relevant and vital. Clear performance goals help a team keep track of progress and hold itself accountable; the broader, even nobler, aspirations in a team's purpose supply both meaning and emotional energy.

Virtually all effective teams we have met, read or heard about, or been members of have ranged between two and 25 people. For example, the Burlington Northern piggybacking team had seven members, and the Knight Ridder newspaper team had 14. The majority of them have numbered less than ten. Small size is admittedly more of a pragmatic guide than an absolute necessity for success. A large number of people, say 50 or more, can theoretically become a team. But groups of such size are more likely to break into subteams rather than function as a single unit.

Why? Large numbers of people have trouble interacting constructively as a group, much less doing real work together. Ten people are far more likely than 50 to work through their individual, functional, and hierarchical differences toward a common plan and to hold themselves jointly accountable for the results.

Large groups also face logistical issues, such as finding enough physical space and time to meet. And they confront more complex constraints, like crowd or herd behaviors, which prevent the intense sharing of viewpoints needed to build a team. As a result, when they try to develop a common purpose, they usually produce only superficial "missions" and well-meaning intentions that cannot be translated into concrete objectives. They tend fairly quickly to reach a point when meetings become a chore, a clear sign that most of the people in the group are uncertain why they have gathered, beyond some notion of getting along better. Anyone who has been through one of these exercises understands how frustrating it can be. This kind of failure tends to foster cynicism, which gets in the way of future team efforts.

For managers to make better decisions about whether, when, or how to encourage and use teams, it is important to be more precise about what a team is and what it isn't.

In addition to finding the right size, teams must develop the right mix of skills; that is, each of the complementary skills necessary to do the team's job. As obvious as it sounds, it is a common failing in potential teams. Skill requirements fall into three fairly self-evident categories.

Technical or Functional Expertise. It would make little sense for a group of

doctors to litigate an employment discrimination case in a court of law. Yet teams of doctors and lawyers often try medical malpractice or personal injury cases. Similarly, product development groups that include only marketers or engineers are less likely to succeed than those with the complementary skills of both.

Problem-Solving and Decision-Making Skills. Teams must be able to identify the problems and opportunities they face, evaluate the options they have for moving forward, and then make necessary trade-offs and decisions about how to proceed. Most teams need some members with these skills to begin with, although many will develop them best on the job.

Interpersonal Skills. Common understanding and purpose cannot arise without effective communication and constructive conflict, which in turn depend on interpersonal skills. These skills include risk taking, helpful criticism, objectivity, active listening, giving the benefit of the doubt, and recognizing the interests and achievements of others.

Obviously, a team cannot get started without some minimum complement of skills, especially technical and functional ones. Still, think about how often you've been part of a team whose members were chosen primarily on the basis of personal compatibility or formal position in the organization, and in which the skill mix of its members wasn't given much thought.

It is equally common to overemphasize skills in team selection. Yet in all the successful teams we've encountered, not one had all the needed skills at the outset. The Burlington Northern team, for example, initially had no members who were skilled marketers despite the fact that their performance challenge was a marketing one. In fact, we discovered that teams are powerful vehicles for developing the skills needed to meet the team's performance challenge. Accordingly, team member selection ought to ride as much on skill potential as on skills already proven.

Effective teams develop strong commitment to a common approach; that is, to how they will work together to accomplish their purpose. Team members must agree on who will do particular jobs, how schedules will be set and adhered to, what skills need to be developed, how continuing membership in the team is to be earned, and how the group will make and modify decisions. This element of commitment is as important to team performance as the team's commitment to its purpose and goals.

Agreeing on the specifics of work and how they fit together to integrate individual skills and advance team performance lies at the heart of shaping a common approach. It is perhaps self-evident that an approach that delegates all the real work to a few members (or staff outsiders) and thus relies on reviews and meetings for its only "work together" aspects, cannot sustain a real team. Every member of a successful team does equivalent amounts of real work; all members, including the team leader, contribute in concrete ways to the team's work product. This is a very important element of the emotional logic that drives team performance.

When individuals approach a team situation, especially in a business setting, each has preexisting job assignments as well as strengths and weaknesses reflecting a variety of talents, backgrounds, personalities, and prejudices. Only through the mutual discovery and understanding of how to apply all its human resources to a common purpose can a team develop and agree on the best approach to achieve its goals. At the heart of such long and, at times, difficult interactions lies a commitment-building process in which the team candidly explores who is best suited to each task as well as how individual roles will come together. In effect, the team establishes a social contract among members that relates to their purpose and guides and obligates how they must work together.

No group ever becomes a team until it can hold itself accountable as a team. Like common purpose and approach, mutual accountability is a stiff test. Think, for example, about the subtle but critical difference between "the boss holds me accountable" and "we hold ourselves accountable." The first case can lead to the second, but without the second, there can be no team.

Companies like Hewlett-Packard and Motorola have an ingrained performance ethic that enables teams to form organically whenever there is a clear performance challenge requiring collective rather than individual effort. In these companies, the factor of mutual accountability is commonplace. "Being in the boat together" is how their performance game is played.

At its core, team accountability is about the sincere promises we make to ourselves and others, promises that underpin two critical aspects of effective teams: commitment and trust. Most of us enter a potential team situation cautiously because ingrained individualism and experience discourage us from putting our fates in the hands of others or accepting responsibility for others. Teams do not succeed by ignoring or wishing away such behavior.

Mutual accountability cannot be coerced any more than people can be made to trust one another. But when a team shares a common purpose, goals, and approach, mutual accountability grows as a natural counterpart. Accountability arises from and reinforces the time, energy, and action invested in figuring out what the team is trying to accomplish and how best to get it done.

A team opportunity exists anywhere hierarchy or organizational boundaries inhibit the skills and perspectives needed for optimal results.

When people work together toward a common objective, trust and commitment follow. Consequently, teams enjoying a strong common purpose and approach inevitably hold themselves responsible, both as individuals and as a team, for the team's performance. This sense of mutual accountability also produces the rich rewards of mutual achievement in which all members share. What we heard over and over from members of effective teams is that they found the experience energizing and motivating in ways that their "normal" jobs never could match.

On the other hand, groups established primarily for the sake of becoming a team or for job enhancement, communication, organizational effectiveness, or excellence rarely become effective teams, as demonstrated by the bad feelings left in many companies after experimenting with quality circles that never translated "quality" into specific goals. Only when appropriate performance goals are set does the process of discussing the goals and the approaches to them give team members a clearer and clearer choice: They can disagree with a goal and the path that the team selects and, in effect, opt out, or they can pitch in and become accountable with and to their teammates.

The discipline of teams we've outlined is critical to the success of all teams. Yet it is also useful to go one step further. Most teams can be classified in one of three ways: teams that recommend things, teams that make or do things, and teams that run things. In our experience, each type faces a characteristic set of challenges.

Teams That Recommend Things. These teams include task forces; project groups; and audit, quality, or safety groups asked to study and solve particular problems. Teams that recommend things almost always have predetermined completion dates. Two critical issues are unique to such teams: getting off to a fast and constructive start and dealing with the ultimate handoff that's required to get recommendations implemented.

The key to the first issue lies in the clarity of the team's charter and the composition of its membership. In addition to wanting to know why and how their efforts are important, task forces need a clear definition of whom management expects to participate and the time commitment required. Management can help by ensuring that the team includes people with the skills and influence necessary for crafting practical recommendations that will carry weight throughout the organization. Moreover, management can help the team get the necessary cooperation by opening doors and dealing with political obstacles.

Missing the handoff is almost always the problem that stymies teams that recommend things. To avoid this, the transfer of responsibility for recommendations to those who must implement them demands top management's time and attention. The more top managers assume that recommendations will "just

happen," the less likely it is that they will. The more involvement task force members have in implementing their recommendations, the more likely they are to get implemented.

To the extent that people outside the task force will have to carry the ball, it is critical to involve them in the process early and often, certainly well before recommendations are finalized. Such involvement may take many forms, including participating in interviews, helping with analyses, contributing and critiquing ideas, and conducting experiments and trials. At a minimum, anyone responsible for implementation should receive a briefing on the task force's purpose, approach, and objectives at the beginning of the effort as well as regular reviews of progress.

Teams That Make or Do Things. These teams include people at or near the front lines who are responsible for doing the basic manufacturing, development, operations, marketing, sales, service, and other value-adding activities of a business. With some exceptions, such as new-product development or process design teams, teams that make or do things tend to have no set completion dates because their activities are ongoing.

In deciding where team performance might have the greatest impact, top management should concentrate on what we call the company's "critical delivery points"–that is, places in the organization where the cost and value of the company's products and services are most directly determined. Such critical delivery points might include where accounts get managed, customer service performed, products designed, and productivity determined. If performance at critical delivery points depends on combining multiple skills, perspectives, and judgments in real time, then the team option is the smartest one.

When an organization does require a significant number of teams at these points, the sheer challenge of maximizing the performance of so many groups will demand a carefully constructed and performance-focused set of management processes. The issue here for top management is how to build the necessary systems and process supports without falling into the trap of appearing to promote teams for their own sake.

The imperative here, returning to our earlier discussion of the basic discipline of teams, is a relentless focus on performance. If management fails to pay persistent attention to the link between teams and performance, the organization becomes convinced that "this year, we are doing 'teams'." Top management can help by instituting processes like pay schemes and training for teams responsive to their real time needs, but more than anything else, top management must make clear and compelling demands on the teams themselves and then pay constant attention to their progress with respect to both team basics and performance results. This means focusing on specific teams and specific performance challenges. Otherwise "performance," like "team," will become a cliché.

Every company faces specific performance challenges for which teams are the most practical and powerful vehicle at top management's disposal.

Teams That Run Things. Despite the fact that many leaders refer to the group reporting to them as a team, few groups really are. And groups that become real teams seldom think of themselves as a team because they are so focused on performance results. Yet the opportunity for such teams includes groups from the top of the enterprise down through the divisional or functional level. Whether it is in charge of thousands of people or just a handful, as long as the group oversees some business, ongoing program, or significant functional activity, it is a team that runs things.

The main issue these teams face is determining whether a real team approach is the right one. Many groups that run things can be more effective as working groups than as teams. The key judgment is whether the sum of individual bests will suffice for the performance challenge at hand or whether the group must deliver substantial incremental performance requiring real joint work products. Although the team option promises greater performance, it also brings more risk, and managers must be brutally honest in assessing the trade-offs.

Members may have to overcome a natural reluctance to trust their fate to others. The price of faking the team approach is high: At best, members get diverted from their individual goals, costs outweigh benefits, and people resent the imposition on their time and priorities. At worst, serious animosities develop that undercut even the potential personal bests of the working-group approach.

Working groups present fewer risks. Effective working groups need little time to shape their purpose, since the leader usually establishes it. Meetings are run against well-prioritized agendas. And decisions are implemented through specific individual assignments and accountabilities. Most of the time, therefore, if performance aspirations can be met through individuals doing their respective jobs well, the working-group approach is more comfortable, less risky, and less disruptive than trying for more elusive team performance levels. Indeed, if there is no performance need for the team approach, efforts spent to improve the effectiveness of the working group make much more sense than floundering around trying to become a team.

Having said that, we believe the extra level of performance teams can achieve is becoming critical for a growing number of companies, especially as they move through major changes during which company performance depends on broad-based behavioral change. When top management uses teams to run

things, it should make sure the team succeeds in identifying specific purposes and goals.

This is a second major issue for teams that run things. Too often, such teams confuse the broad mission of the total organization with the specific purpose of their small group at the top. The discipline of teams tells us that for a real team to form, there must be a team purpose that is distinctive and specific to the small group and that requires its members to roll up their sleeves and accomplish something beyond individual end products. If a group of managers looks only at the economic performance of the part of the organization it runs to assess overall effectiveness, the group will not have any team performance goals of its own.

While the basic discipline of teams does not differ for them, teams at the top are certainly the most difficult. The complexities of long-term challenges, heavy demands on executive time, and the deep-seated individualism of senior people conspire against teams at the top. At the same time, teams at the top are the most powerful. At first we thought such teams were nearly impossible. That is because we were looking at the teams as defined by the formal organizational structure; that is, the leader and all his or her direct reports equals the team. Then we discovered that real teams at the top were often smaller and less formalized: Whitehead and Weinberg at Goldman Sachs; Hewlett and Packard at HP; Krasnoff, Pall, and Hardy at Pall Corporation; Kendall, Pearson, and Calloway at Pepsi; Haas and Haas at Levi Strauss; Batten and Ridder at Knight Ridder. They were mostly twos and threes, with an occasional fourth.

Nonetheless, real teams at the top of large, complex organizations are still few and far between. Far too many groups at the top of large corporations needlessly constrain themselves from achieving real team levels of performance because they assume that all direct reports must be on the team, that team goals must be identical to corporate goals, that the team members' positions rather than skills determine their respective roles, that a team must be a team all the time, and that the team leader is above doing real work.

As understandable as these assumptions may be, most of them are unwarranted. They do not apply to the teams at the top we have observed, and when replaced with more realistic and flexible assumptions that permit the team discipline to be applied, real team performance at the top can and does occur. Moreover, as more and more companies are confronted with the need to manage major change across their organizations, we will see more real teams at the top.

We believe that teams will become the primary unit of performance in high-performance organizations. But that does not mean that teams will crowd out individual opportunity or formal hierarchy and process. Rather, teams will enhance existing structures without replacing them. A team opportunity exists anywhere hierarchy or organizational boundaries inhibit the skills and perspectives needed for optimal results. Thus, new-product innovation requires preserving functional excellence through structure while eradicating functional bias through teams. And frontline productivity requires preserving direction and guidance through hierarchy while drawing on energy and flexibility through self-managing teams.

We are convinced that every company faces specific performance challenges for which teams are the most practical and powerful vehicle at top management's disposal. The critical role for senior managers, therefore, is to worry about company performance and the kinds of teams that can deliver it. This means top management must recognize a team's unique potential to deliver results, deploy teams strategically when they are the best tool for the job, and foster the basic discipline of teams that will make them effective. By doing so, top management creates the kind of environment that enables team as well as individual and organizational performance.

Reprint R0507P; HBR OnPoint 4428
To order, see page 195.

Leading High-Impact Teams
september 25–28

Biotechnology: Science for Managers
october 2–5

Biotechnology: Strategies for Growth
october 2–5

Business Marketing Strategy
october 2–7

Managing Customer Relationships for Profit
october 2–5

The Soul of Leadership
october 5–7

Kellogg on Branding: Building, Leveraging and Rejuvenating Your Brand
october 9–12

Reinventing Leadership: A Breakthrough Approach
october 9–14

Upcoming programs where the business world's most courageous minds tackle its most challenging issues.
execed.kellogg.northwestern.edu

847-491-3100

Kellogg
School of Management

Harvard Business Review Notice of Use Restrictions, May 2009

Harvard Business Review and Harvard Business Publishing Newsletter content on EBSCOhost is licensed for the private individual use of authorized EBSCOhost users. It is not intended for use as assigned course material in academic institutions nor as corporate learning or training materials in businesses. Academic licensees may not use this content in electronic reserves, electronic course packs, persistent linking from syllabi or by any other means of incorporating the content into course resources. Business licensees may not host this content on learning management systems or use persistent linking or other means to incorporate the content into learning management systems. Harvard Business Publishing will be pleased to grant permission to make this content available through such means. For rates and permission, contact permissions@harvardbusiness.org.

DHQ: Digital Humanities Quarterly

2009
Volume 3 Number 2

Done: Finishing Projects in the Digital Humanities

Matthew G. Kirschenbaum <mgk_at_umd_dot_edu>, University of Maryland

Abstract

How do we know when we're done? This cluster of articles explores completion and incompletion in the digital humanities from a variety of perspectives.

Done. Way back when I worked as the project manager for the *William Blake Archive*, there was no more satisfying sequence of characters than those five keystrokes (including punctuation). *Record newly received transparencies in project log.* Done. *Parse files in holding directory.* Done. *Purchase new scanner.* Done. *Proofread documentation for image annotation applet.* Done. It was easy to measure progress as I ticked off tasks in email messages to the archive's editors. In time, when enough of these individual tasks were "done," the project might be finished. 1

Or would it? How do we *know* when we're done? What does it mean to "finish" a piece of digital work? As Bill Kretzschmar points out in his essay, the verb "to finish" can mean to *complete* or something more like to polish or perfect. What is the measure of "completeness" in a medium where the prevailing wisdom is to celebrate the incomplete, the open-ended, and the extensible? Digital humanities and the electronic publishing ventures associated with it are used to deriving considerable rhetorical mileage and the occasional moral high-ground by contrasting their radical flexibility and mutability with the glacial nature of scholarly communication in the fixed and frozen world of print-based publication. 2

It is also, let us admit, usually more exciting to begin something than to finish it. We partake in what Julia Flanders has aptly called the culture of the perpetual prototype: the demo, the proof-of-concept, the alpha and beta version. Building things is fun, for a lot of us it's why we got into this business; debugging and documenting them, tweaking and patching them is not. This tendency extends to our grant cycles as well. In the United States, we now have the great benefit of digital humanities start-up grants, but not dedicated awards for finishing or closing projects. Granting agencies reward innovation and originality, not steady, measured progress. No one writes a grant application to say, "With just two more years of funding we can finish the good work we've started. No, we're not going to be doing anything new. We just want finally to put this thing to bed." 3

It may be tempting to write off worn out tools, abandoned projects, and occasionally even outright vaporware as the victims of a rapacious technological obsolescence stalking all our endeavors. But this doesn't tell the whole story. Nor is the digital humanities unique in its uncertainty — arguably its timidity — over knowing when to call something done. As Silicon Valley journalist Scott Rosenberg suggests, software development has an important analog in a specific point of computational theory, the "halting problem" as defined by Alan Turing. Knowing how or when a particular version or release is finished and ready for public debut is an ongoing and often heated debate among coders. Rosenberg quotes from Alan Cooper's *The Inmates are Running the Asylum*: "Software development lacks one key element — an understanding of what it means to be 'Done'" [Rosenberg 2007]. 4

But we cannot escape the pressure of milestones, deadlines, deliverables, and products (as opposed to projects) as the measure of academic success and the currency of professional promotion and reward. They are also what allow our work to escape a small circle of developers and enthusiasts and become part of a larger scholarly community, where it can be used in the consequential manner presumably always intended. 5

At Digital Humanities 2007 I organized a panel to begin addressing what it actually means to finish work that we've started. The essays that follow are significantly revised and expanded from the remarks the authors presented at the conference. William A. Kretzschmar's piece traces the local history of one particular project, grafting milestones and markers of what's been "done" to various versions and releases as his tools and technologies evolve. His remarks raise issues not only for how one assesses closure, but also digital preservation and version control. David Sewell addresses the questions I posed from the perspective of a publisher. His work at the University of Virginia Press's Rotunda digital imprint offers a different context, one where finding some measure of "done" is a professional and financial necessity. Finally, Susan Brown and her colleagues from the *Orlando* project offer the view from one of the largest and longest running undertakings in the digital humanities, at the moment when the project has gone to press. Brown discusses what got done and what it will take for it to stay that way. 6

Works Cited

Rosenberg 2007 Rosenberg, Scott. *Dreaming in Code*. New York: Crown, 2007, 337.

DHQ: Digital Humanities Quarterly

2009
Volume 3 Number 2

Large-Scale Humanities Computing Projects: Snakes Eating Tails, or Every End is a New Beginning?

William A. Kretzschmar, Jr. <kretzsch_at_uga_dot_edu>, University of Georgia

Abstract

The word "finish" can mean two things that have quite different implications for large-scale humanities computing projects: "to bring to completion; to make or perform completely; to complete" and also "to perfect finally or in detail; to put the final and completing touches to (a thing)." The word "finish" is just not part of the deal for the Linguistic Atlas Project in either sense. However, granting agencies must ask "what do you want money for this time?" and, from this viewpoint, the Atlas Project consists of a series of particular tasks or experiments, each one of which is capable of being "finished" in both senses of the word. This paper discusses the reality of funding, deadlines, and deliverables, as they relate to the sequence of tasks that make up the larger Atlas Project. There are no once-and-done, permanent solutions. The largest humanities computing projects require continuing care and maintenance, and the best way forward is to create some sort of stable institutional setting for large projects that will provide continuity and baseline resources for the work.

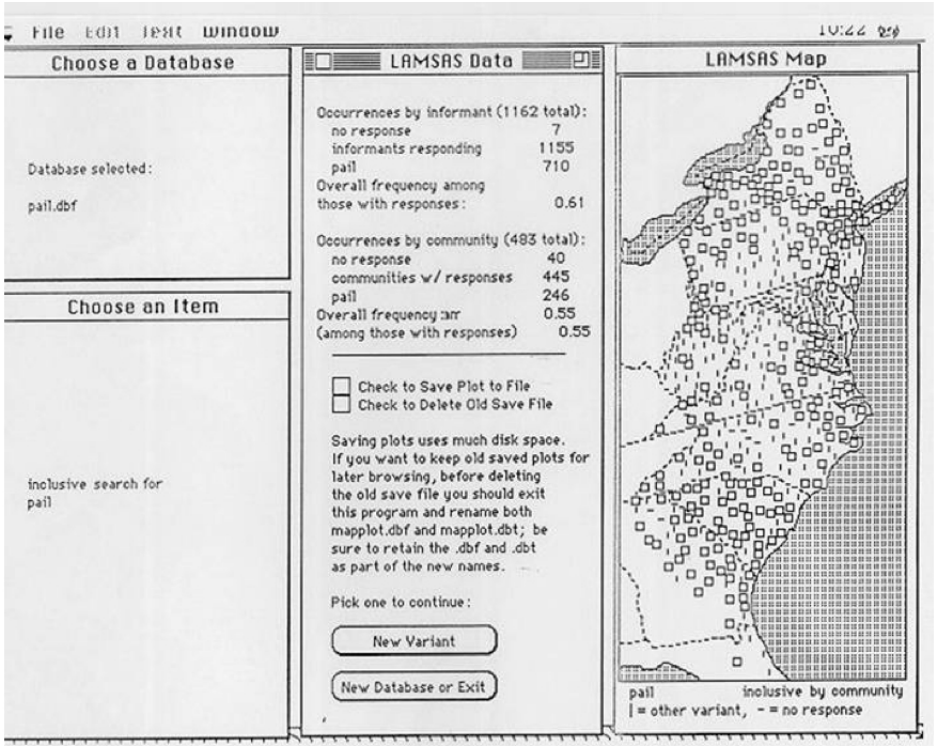
One of the motivating questions for this cluster of essays is "What does it mean to finish something?" As it happens, the word "finish" can mean two things that have quite different implications for large-scale humanities computing projects. On one hand, according to the *OED*, "finish" can mean "To bring to completion; to make or perform completely; to complete." On the other hand, the word can also mean "To perfect finally or in detail; to put the final and completing touches to (a thing)." In my own work of this kind, the American Linguistic Atlas project (http://www.lap.uga.edu), we do neither of these things. We cannot come to an end of the work because we are witnesses and archivists of how Americans talk, and they keep talking differently across time and space. Neither do I think that our humanities-computing representation of our research is capable of being finally perfected, of achieving some perfect state, because technology keeps changing and the demands placed upon our research keep changing. If we view the entirety of the Linguistic Atlas Project as a "large-scale humanities computing project," the word "finish" is just not part of the deal. And we are not alone. While the creation of, say, a variorum edition may seem like a project that can be finished in both senses, actually we need to make new editions all the time, since our idea of how to make the best edition changes as trends in scholarship change, especially now in the digital age when new technical possibilities keep emerging. 1

However, it is quite reasonable to ask, as our granting agencies must ask, "What do you want money for this time?" or "Did you accomplish what we gave you money to do?" From this viewpoint, the Atlas Project, as an example to stand in for any large scale project, consists of a series of particular tasks or experiments, each one of which is capable of being "finished" in both senses of the word. It is these separate tasks that are fundable, and for which we can claim to have done what we said we would do. In this paper, I would like to discuss the sequence of tasks that make up the larger Atlas Project in order to show the special character of work done deliberately as part of a sequence for a large-scale project, as opposed to work proposed as a singular task. 2

The contextualization of the separate tasks leads to special cases of what it means to "finish" the work in either sense. The point of what follows is not the Atlas Project itself, but instead the way that individual tasks respond to the technical and academic situation at the time. Our own technical work on the Atlas has responded to the microcomputer revolution, to the emergence of the Web, and to the development of text-encoding and computer multimedia. Our work has helped to drive changes in the academic study of language variation, from traditional dialect maps to use of rapid visualization methods and statistical processing for both text and audio data. The scope of these changes show how our work and thinking over the years have had to change and must continue to change, just as they must for other large-scale projects like major editions and dictionaries, so that we can avoid the charge of being the snake that eats its own tail.

The Linguistic Atlas Project, per se, has been notable over many years for its twin goals of interactivity for research (including the use of GIS) and making its data sets accessible and available to other researchers and to the public. I first programmed a GIS system for our Linguistic Atlas data in 1990, presented at the ACH/ALLC meeting in Tucson in 1991 Figure 1; [Kirk and Kretzschmar 1991]; [Kirk and Kretzschmar 1992]; [Kretzschmar 1992]. The program, called LAMSASplot, took advantage of work we did with funding from NEH to keyboard words and phrases from our survey interviews so that they would be available for computer-assisted analysis. The LAMSASplot system was widely used for teaching and research on American English in the early 1990s, and it immediately led to breakthroughs in how we were able to think about language variation data [Kretzschmar 1994], [Kretzschmar 1996a]; [Lee and Kretzschmar 1993]. The two most important design elements of that first system were the central column, which shows the frequency of occurrence of a word of interest, both by speaker and by community (most communities have more than one speaker), and the rightmost column, the GIS implementation. That column is composed of two layers, a standard base map layer that shows the state outlines, and a second layer consisting of plotted points at community locations that is generated according to the evidence for the word selected.

Figure 1.

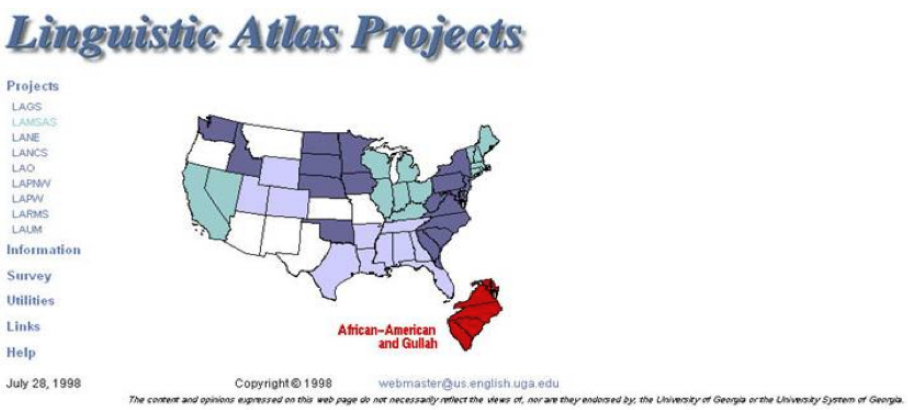


LAMSASplot on the Mac platform, c. 1990

The LAMSAS GIS programs were prepared on Macintosh computers using the Foxbase relational database package. Foxbase permitted us to associate graphical coordinates with our linguistic data within a relational structure. Generation of these plots took only 90 seconds on the desktop computers of the day, at least a 250-fold improvement over the hours required for the charting that had to be done by hand by the pioneers in the field, Hans Kurath and Raven McDavid. The other advantage of LAMSASplot is that it made frequency counts and plotted one variant at a time. This practice was in sharp contrast to the method used by Kurath and McDavid for charting language data, using isoglosses, lines which showed the limit of occurrence of particular variants. As our pail map suggests here, lines just cannot do justice to the geographical distributions of linguistic features. Our GIS solution thus launched a new line of analysis for dialectologists based on differential frequency in feature distributions. The LAMSASplot humanities computing application not only used available technology to automate mapmaking, it provided a theoretical advance in how the data was analyzed. The Atlas project was thus one of many involved in the burgeoning field of computer-assisted scientific visualization, not just in the humanities, but the particular point of talking about the Atlas is to show how such large trends have particular effects in particular areas of study. 4

We did not finish keyboarding all of our LAMSAS data with our NEH funding, which had been proposed as the first of a sequence of grants. We could show immediate benefits of getting the process started, in part through LAMSASplot, but unfortunately NEH did not fund our proposals to continue digitization of the data. There were various reasons offered by the panelists over the years, but I suspect that the refusals came down to the fact that the panelists preferred to fund new work over our historical interviews. The immediate analytical task for the Atlas was thus "finished" in both senses, both completed and refined, but only the immediate task; we were shut out from long-term funding to "finish" the whole project. 5

Figure 2.

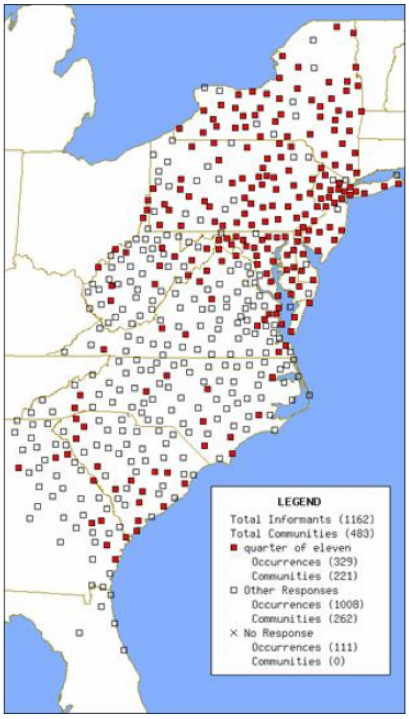


Atlas Web Site, Mark 1

As the next task in our humanities computing implementation, we then ported the system to the Web, which I first demonstrated in 1996 ([Kretzschmar 1996b]; Figure 2 shows the site as of 1998). We had been working on an interactive ftp/gopher system as early as 1993, but when Web technology became available we saw that it enabled perfectly what we had been attempting from another direction. The Web allowed us to make our data available to a wide public, with many additional interactive GIS features for locating speakers and information for the LAMSAS survey, and with static pages for the other regional surveys. We added additional data for our African American LAMSAS speakers plus data from speakers of Sea Island Creole (Gullah) in 1998 with 6

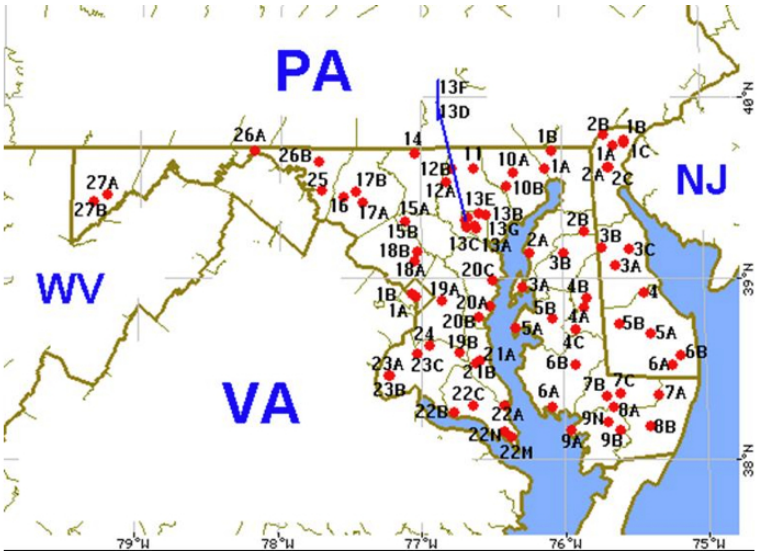
funding from NSF. The GIS maps from LAMSASplot were ported to the Web, as in Figure 3 for the phrase *quarter of* in telling time (as opposed to *quarter till*, *quarter to*), where they now could be created by users almost instantaneously with a server-side script, another improvement in processing time of perhaps fiftyfold. We also offered maps to show where all of the speakers in a state resided (Figure 4). These maps were clickable to allow users to access all of the information about any given speaker and the speaker's survey responses (Figure 5). Besides maps, we enabled browsing and searches of the responses and of speaker information (Figure 6, Figure 7). We also changed the way that we stored our data: flat, comma-delimited files instead of a proprietary relational format.

Figure 3.



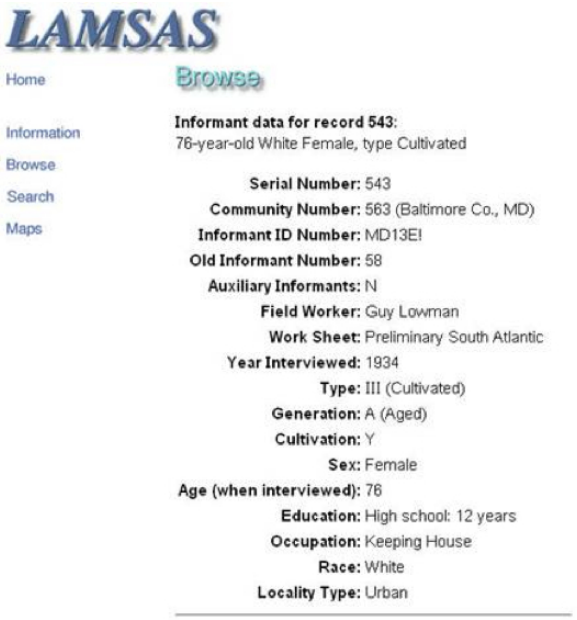
LAMSAS Web: *quarter of*

Figure 4.



LAMSAS Web: state map of speakers in Maryland

Figure 5.



LAMSAS Web: Baltimore speaker (from clicking MD13E

Figure 6.

LAMSAS Web: Browse Screen

Figure 7.

LAMSAS Informant Database Plain Format

Links:

- Explanation of the fields
- Search the informants
- Examine the raw data file
- Back to Browse

Inf#	Typ	Community	Local	WS	Year	Sex	Age	Race	Gen	Cult	Educ	Occup
NY1	I	Suffolk Co	R	E	1933	M	53	W	B	N	3	F
NY2A	I	Suffolk Co	R	E	1933	M	63	W	A	N	O	F
NY2B	I	Suffolk Co	R	E	1933	M	85	W	A	N	3	F
NY2C	II	Suffolk Co	U	E	1933	F	80	W	A	N	2	K
NY3A	I	Suffolk Co	R	M	1941	M	71	W	A	N	4	F
NY3B	I	Suffolk Co	R	M	1941	M	70	W	A	N	3	F
NY3C	I	Suffolk Co	R	M	1941	M	50	W	B	N	4	F
NY3D	I	Suffolk Co	R	M	1941	M	48	W	B	N	3	F
NY4A	I	Nassau Co.	R	M	1941	M	69	W	A	N	2	F
NY4B	I	Nassau Co.	U	M	1941	M	79	W	A	N	2	O
NY4C	I	Nassau Co.	R	M	1941	M	51	W	B	N	3	F
NY5A	I	Brooklyn	U	M	1941	F	70	W	B	N	4	K
NY5B	I	Brooklyn	U	M	1941	M	62	W	B	N	4	R
NY5C	II	Brooklyn	U	M	1941	F	51	W	B	N	3	K
NY5D	II	Brooklyn	U	M	1941	M	64	W	B	N	2	R
NY5E	II	Brooklyn	U	M	1941	F	82	W	A	N	2	K
NY5F	II	Brooklyn	U	M	1941	M	53	W	B	N	2	M
NY5G	II	Brooklyn	U	M	1941	M	50	W	B	N	4	M
NY5H!	III	Brooklyn	U	M	1941	M	68	W	A	Y	6	P
NY5I!	III	Brooklyn	U	M	1941	M	54	W	B	Y	6	M
NY6A	I	Richmond C	U	M	1941	F	76	W	A	N	2	K
NY6B	II	Richmond C	U	M	1941	F	48	W	B	N	4	K
NY7A	I	Manhattan	U	M	1941	M	45	W	B	N	2	L
NY7B	I	Manhattan	U	M	1941	M	39	W	B	N	2	O

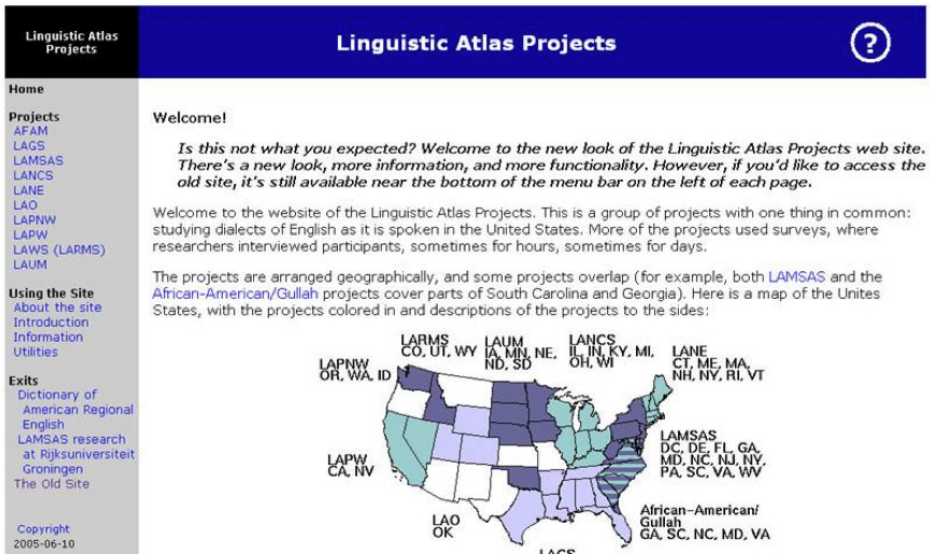
LAMSAS Web: Speakers (click of Figure 6, plain format informants

The Atlas Web was a significant advance for both teaching and research ([Kretzschmar 1997], [Kretzschmar 2002a]), in line with the goals for an electronic atlas first set forth nearly a decade earlier [Kretzschmar 1988]. The plotted maps and much of the data were the same as what was available for the Mac LAMSASplot package — besides the African American data, we had keyboarded some additional files, a few at a time as resources permit, as we are still doing — but instead of an incremental improvement, the Atlas Web site was 7

actually another sea change for the project. For the first time the whole Atlas project was represented together, not just as separate regional parts, and for the first time we could offer real public access to as much data as we could digitize. Again, the example of the Atlas just shows the particular significance in our area of our participation in the main international trend for Web representation of information.

These were good things, revolutionary in their way, but the fact is that we *had* to do something: the world had changed from mainframes and desktops to UNIX servers. We had actually been using both mainframe processing (for intensive statistics jobs) as well as the desktop Macs and PCs. If we had not acted, the proprietary software that we were using would have gone out of use (FoxBase was sold, and the database for our PCs, RBase, also disappeared), and so we would have lost our investment in preparation of proprietary data files. We would have missed the Web revolution. So, not to change would have meant that the Atlas project would be stuck on paper, where it had been before we developed any computer applications at all. We did benefit from the NSF grant I mentioned, and from a small internal University of Georgia grant that allowed us to set up our own server, both limited tasks that we could propose as achievable within a short time and for which we could show results. We still could not get funding to do the whole project, and if we had waited for that we would still be waiting. Every large humanities computing project faced similar problems: adapt, or become a footnote in academic history. 8

Figure 8.



LAMSAS Web, Mark 2, c. 2005

After a while, we wanted to do new things, so we began work on a major revision of the Web site that came on line in 2003 (Figure 8, shown as of 2005). Our earlier GIS visualizations were just not as flexible as they needed to be to satisfy the demands of sociolinguistics, which had come to expect association of linguistic features with particular social variables (e.g. sex, age, social status). We kept the interactive plotted maps and layered GIS access to information, but added even more interactive choices such as more flexible searches and tallies of the speakers and language data set. Browse screens are now composed on the fly from separate data files, as in Figure 9 from the quarter of data and speaker characteristics, so that the linguistic information is immediately associated with the social characteristics of those who said it. For the first time we enabled searching by pronunciation features, not just words or strings, and we offered the ability to do sub-searches in sequence so that users could zero-in on features and groups of interest (Figure 10). In order to accomplish these goals, we again changed our manner of storing and manipulating data from a homemade phonetic font to 9

Unicode, and from flat files and scripted regular expressions to a MySQL implementation. However, we kept "The Old Site" as a link on the new one, so that long-time users would find what was familiar to them, and also for users who did not want the greater complexity that came with greater flexibility of use (see the link on the bottom left corner of Figure 8). We could not just move it, however, because "The Old Site" had to be compatible with Unicode and with the extensive Python scripting that ran functions on the New Site. The task of importing the Mac-based GIS system to the Web was complete by 1996, but was finished, in that second sense, in 2003 with the platform change and the new touches of the more flexible site. Again, other large-scale humanities computing projects faced similar problems; we all had to adapt to new scripting languages and environments, new tools like MySQL and Unicode, and new scholarly trends and demands in our particular areas.

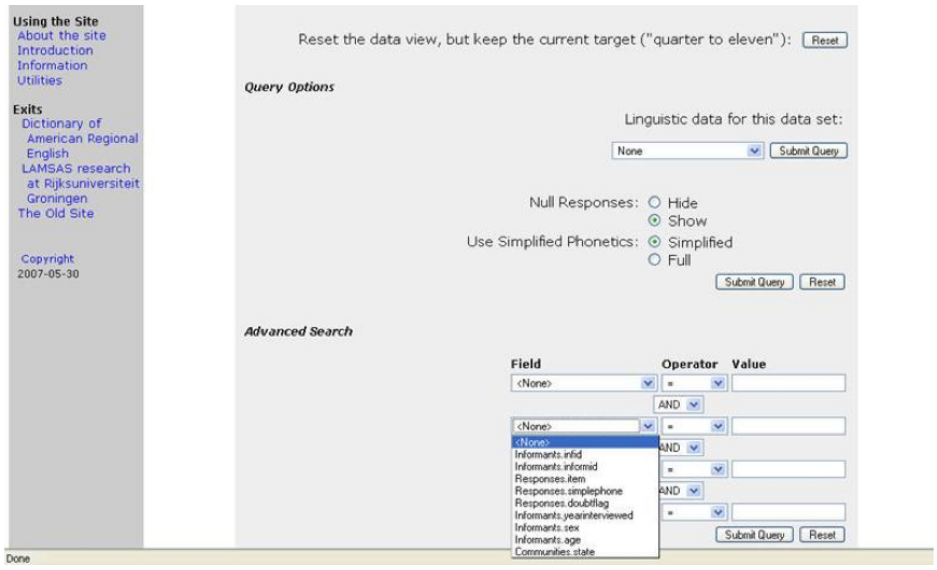
10

The transition to the Web and the further refinement of our Web tools in the late 1990s and early years of the new century were not accomplished by means of external funding from federal sources. We did win awards to record the African American data and to conduct new kinds of research (e.g., NSF awards to explore neural-network analysis of Atlas data, and to conduct a pilot project for random-sample survey research in Atlanta), and some of these resources were applied to computer work, but no award specifically funded our Web innovations. Development of our Web site was primarily accomplished by two graduate assistants, Rafal Konopka and Eric Rochester, who worked on Atlas computer projects for twelve years between them, first Konopka through our initial Web implementation, then Rochester through the first major revision. They held graduate assistantships, sometimes fully funded for research but more often cobbled together from different sources including both research and teaching. The consistent, stable element in this funding came from the Hans Kurath Fund of the American Dialect Society. The Kurath Fund is an endowment for the Atlas created by Raven McDavid and maintained by the American Dialect Society [Kretzschmar 2003]. The Dialect Society hold title to the Atlas research materials, and its Executive Council ratifies editors and advisory board members for the project. While the Kurath Fund could not support a complete graduate assistantship, it could pay enough so that other funding could round out a position. Similarly, while the Kurath Fund could not pay all of the operational expenses of the project, it could provide a key piece of equipment from time to time, repairs, or specialized supplies. The University of Georgia had agreed to provide space for the project and some operational support, in conjunction with the author's faculty appointment (a more permanent agreement may in time be possible). Thus the funding that allowed continuous development of the Atlas Web site has come from multiple sources, and still does. We are always in the position of putting together the pieces of the funding structure year by year. The central role of the American Dialect Society, however, provides the essential stability that has made continuous operation of the editorial site possible.

Figure 9.

LAMSAS Web 2: Browse Screen

Figure 10.



LAMSAS Web: Sub-Search

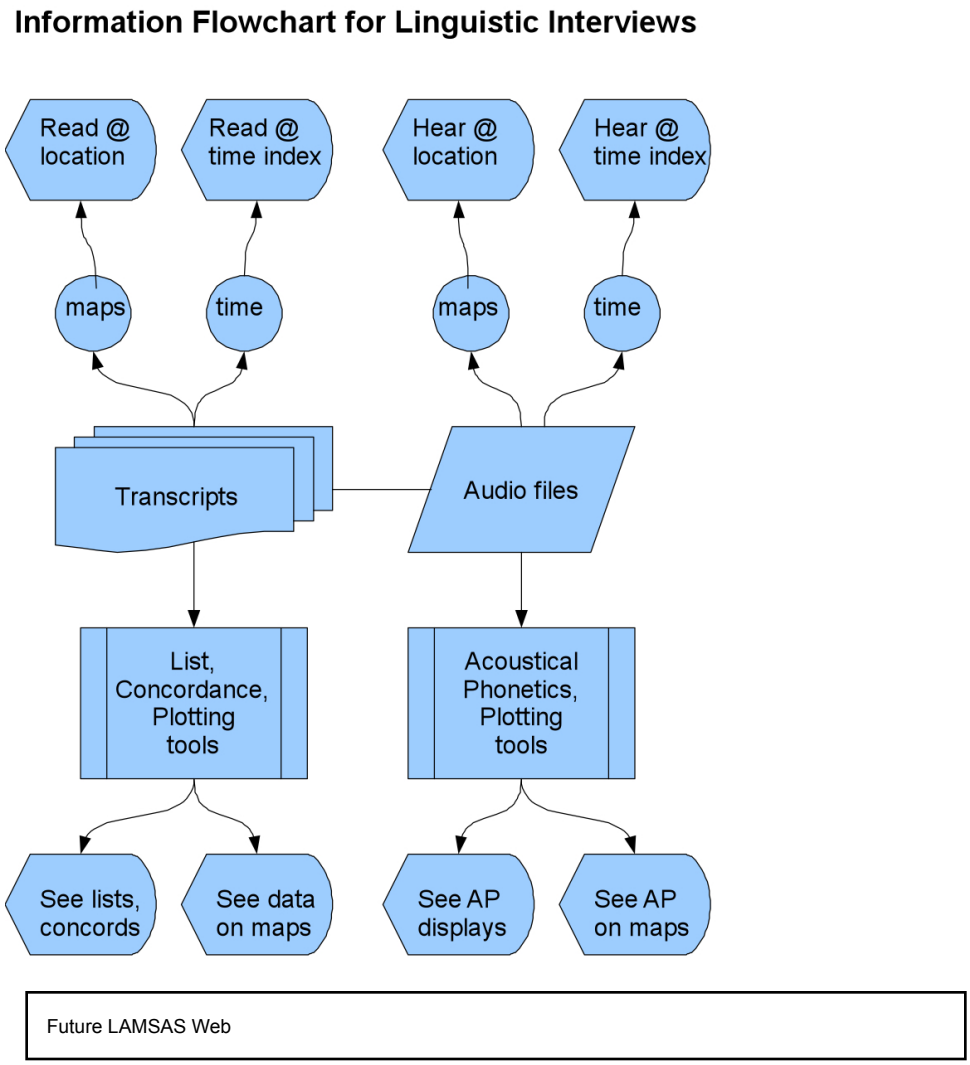
Figure 11.

LAMSAS Web: Work by Others

The additional changes for the LAMSAS Web Mark 2 may seem more incremental than revolutionary, but again we feel that they responded to a changing environment that could not be ignored, new scholarly imperatives. For the first time our linguistic survey data was now fully available for sociolinguistic research. Dialectology is a venerable pursuit in linguistics, but it had been overtaken by sociolinguistics in recent decades and declared essentially to be irrelevant by many sociolinguists. The changes that we made to our Web site integrated social and geographical analysis of language variation in a single visualization that, in itself, showed graphically that the two approaches were not only compatible but inseparable. In addition, we also began to link our site with other sites, so that the Atlas Project could be seen in its connections with other online resources. Finally, we began to post completed analyses on the site in addition to raw data and associated information; some of these are results from our own funded research, others, such as those from John Nerbonne in Groningen (Figure 11), are from the research of people who used our freely-available data. In this way we could make and defend the claim that our Atlas survey research was a member of a wider community of those studying language variation. We did respond to emerging computational opportunities in the New Site, like Unicode and Python and MySQL, but the most significant part of the new changes spoke to issues of linguistic theory and analysis. And again, nobody funded completion of digitization of all of the data from the Atlas.

11

Figure 12.



Still, the larger Atlas Project is nowhere near at an end. We are now rethinking what the site should do, from a text-based system to one that features audio and stored images along with text and GIS. This change has become possible only in the last two years, as much greater network-attached storage has become available (measured in Tb, before long Pb). Because of our archival audio files, we are now one of the largest clients at the University of Georgia institutional storage array (which we share with bio-informaticists, physicists, and others usually considered to be power users). We now conceive of our new interviews as conversational corpora, in which text transcriptions serve as time-linked indices to audio files [Kretzschmar 2005a]; [Kretzschmar 2006]. Figure 12 charts the flow of information that we envision for the next-generation site. While many users will want to listen to our speakers, others will want to perform acoustical analyses, now a strong trend in language variation research, as we ourselves now perform them (e.g., [Kretzschmar 2004]; [Kretzschmar 2005b]). Our next task is to integrate sound with text and to enable acoustical research functions, while maintaining our interactive GIS functions — a whole new set of tools and problems from the previous task [Kretzschmar 2002b]. As the flowchart suggests, we now envision eight different outputs from a new Web presence: full text transcripts, linked sound, acoustical phonetic data in lists and plots, tallies of feature variants in lists and on maps, and technical statistical results in both lists and maps. While any single task from this list is possible, some will be difficult to automate, and the hardest thing of all will be to integrate all eight outputs. We have not yet made these changes; they have been designed for some time now, and await the right people and circumstances and resources to make them real (our major revision cycle seems to be 5 to 10 years).

12

When we can do that, we will have integrated our existing modes of study for language variation, surveys and sociolinguistics, with corpus linguistics, the latest major contemporary approach to language variation. We will have provided not just words and phrases extracted from interviews, but a full complement of information in our interviews, from sound recording to transcripts to analyses of acoustical characteristics.

13 So, will we ever be "finished"? Yes, with the GIS of LAMSASplot and of "The Old Site," though still (as always) tweaking the current site. We have plans to make yet another thorough revision. But we will never be finished either with keeping older sites and available or with creating new visualizations for the information we keep, as new technical possibilities and research demands appear. We can complete particular tasks, and often we can even "finish" particular tasks in the sense of polishing them for improved use. Yet one research proposal does not make the whole research program. While we can often fund and succeed with individual tasks, we must always see tasks as part of the larger process that probably will not be funded completely but still must continue for future users. We would be the snake eating its own tail if all we did was keep polishing eternally a single task (or worse, the dog that pointlessly chases his own tail), but we do well to make every end a new beginning as new technological possibilities become available, and new theories and styles of analyses take hold.

14 After twenty-five years of trying to apply humanities computing to the problems of the Linguistic Atlas, it has become clear that there are no once-and-done, permanent solutions. The largest humanities computing projects are likely to require continuing care and maintenance, if not more radical representation and reinterpretation in light of the advance of scholarship, and yet they seem unlikely ever to be funded comprehensively for these tasks. The best way forward is to create some sort of stable institutional setting for large projects that will provide continuity and baseline resources for the work. This we have done for the Linguistic Atlas, through the American Dialect Society and our association with the University of Georgia. Stable institutional settings allow for additional resources to be sought for individual tasks, as the need for each one becomes apparent. In my general area, the Dictionary of American Regional English project at Wisconsin and the Dictionary of Old English project at Toronto, both innovative users of humanities computing, have found long-term funding through NEH and have also had strong institutional support. We can hope that Wisconsin and Toronto will not end their association with these projects when their first editions are completed. A good model for what might happen has occurred at Michigan, where the completion of the (print) *Middle English Dictionary* was followed by the creation of the electronic *Middle English Compendium* [McSparran 2000] in conjunction with the University of Michigan Library. Conversion was assisted by a grant from NEH. We are now attempting something of the kind at Georgia, with our new grant from NEH which will allow us to make digital conversions of the existing audio-taped interviews from the Linguistic Atlas, and which will also help us to make a new public audio archive of these materials.

15 Thus, at the end of the day, I recommend that, besides grants for new tasks and new work, granting agencies should consider assisting in the formation of enduring institutional structures to support large-scale projects in humanities computing. Some existing programs, like NEH Challenge Grants, may already be helpful for this purpose. The establishment of common interests between the library community and researchers in humanities computing, already well started in the ACH and ALLC and exemplified by projects like the *Middle English Compendium*, is an excellent trend to continue. We will probably never be able to fund big projects like mine at the level needed to finish them completely. But that is OK, because the idea of finishing big humanities computing projects once and for all is just an illusion. We will always need to address changes in technology and changes in our disciplines, even for subject materials that are historical and unchanging. We need, therefore, to accept the rule for Talmudic scholarship that my colleague Virginia McDavid once passed along to me: "You do not have to finish the work, but neither may you desist from it." We just have to keep at it, and find the resources we need along the way.

Works Cited

Kirk and Kretzschmar 1991 Kirk, John, and William A. Kretzschmar, Jr. *The Analysis and Interpretation of Dialect Databases by Interactive Mapping.* ACH/ALLC Conference, Tempe, 1991.

Kirk and Kretzschmar 1992 Kirk, John, and William A. Kretzschmar, Jr. 1992. "Interactive Linguistic Mapping of Dialect Features". *Literary and Linguistic Computing* 7: 168-75.

Kretzschmar 1988 Kretzschmar, William A. Jr. 1988. "Computers and the American Linguistic Atlas". In *Methods in Dialectology: Proceedings of the Sixth International Conference on Methods in Dialectology*, edited by Alan Thomas, 200-24. Clevedon: Multilingual Matters.

Kretzschmar 1992 Kretzschmar, William A. Jr. 1992. "Interactive Computer Mapping for the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS)". In *Old English and New: Essays in Language and Linguistics in Honor of Frederic G. Cassidy*, edited by N. Doane, J. Hall, and R. Ringler, 400-14. New York: Garland.

Kretzschmar 1994 Kretzschmar, William A. Jr. 1994. "Linguistic Theory and Computer Modeling of Linguistic Survey Data." ACH/ALLC, Paris, 1994.

Kretzschmar 1996a Kretzschmar, William A. Jr. 1996a. "Quantitative Areal Analysis of Dialect Features". *Language Variation and Change* 8: 13-39.

Kretzschmar 1996b Kretzschmar, William A. Jr. 1996b. The LAMSAS Internet Site, NWAVE, Las Vegas.

Kretzschmar 1997 Kretzschmar, William A. Jr. 1997. "Computer-Assisted Study of American English Lexical Data". In *From AElfric to the New York Times: Studies in English Corpus Linguistics*, edited by Udo Fries, Viviane Müller, and Peter Schneider, 239-47. Amsterdam, Atlanta: Rodopi.

Kretzschmar 2002a Kretzschmar, William A. Jr. 2002a. "Teaching American English Online". *Journal of English Linguistics* 30: 318-327.

Kretzschmar 2002b Kretzschmar, William A. Jr. 2002b. TEI and Linguistic Interviews. http://www.tei-c.org/.

Kretzschmar 2003 Kretzschmar, William A. Jr. 2003. "Linguistic Atlases of the US and Canada". In *Needed Research in American Dialects*, ed. by Dennis Preston. *Publications of the American Dialect Society* 88: 25-48.

Kretzschmar 2004 Kretzschmar, William A., Jr., Sonja Lanehart, Betsy Barry, Iyabo Osiapem, and MiRan Kim. 2004. Atlanta in Black and White: A New Random Sample of Urban Speech. NWAVE 2004, Ann Arbor.

Kretzschmar 2005a Kretzschmar, William A. Jr., Betsy Barry, and Nicole Kong. 2005. Publication of Full Interviews from the Atlanta Survey Project. ADS/LSA 2005, Oakland.

Kretzschmar 2005b Kretzschmar, William A. Jr., MiRan Kim, and Nicole Kong. 2005. Vowel Formant Characteristics from the Atlanta Survey Project. ADS/LSA 2005, Oakland.

Kretzschmar 2006 Kretzschmar, William A. Jr., Jean Anderson, Joan Beal, Karen Corrigan, Lisa-Lena Opas-Hanninen, and Bartek Plichta. 2006. "Collaboration on Corpora for Regional and Social Analysis". *Journal of English Linguistics* 34: 172-205.

Lee and Kretzschmar 1993 Lee, Jay, and William A. Kretzschmar Jr. 1993. "Spatial Analysis of Linguistic Data with GIS Functions". *International Journal of Geographical Information Systems* 7: 541-60.

McSparran 2000 McSparran, Frances. 2000. *Middle English Compendium*. http://quod.lib.umich.edu/m/mec/ 1954-2007. *Middle English Dictionary*. Ed. by Hans Kurath, Sherman Kuhn, Robert Lewis, et al. Ann Arbor: University of Michigan Press.

NATIONAL INITIATIVE FOR A NETWORKED CULTURAL HERITAGE
ninch

search

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

By the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and the National Initiative for a Networked Cultural Heritage (NINCH)

Acknowledgements

I. Introduction

II. Project Planning

III. Selecting Materials: An Iterative Process

IV. Rights Management

V. Digitization and Encoding of Text

VI. Capture and Management of Images

VII. Audio/Video Capture and Management

VIII. Quality Control and Assurance

IX. Working With Others

X. Distribution

XI. Sustainability: Models for Long-Term Funding

XII. Assessment of Projects by User Evaluation

XIII. Digital Asset Management

XIV. Preservation

Appendix A: Equipment

Appendix B: Metadata

Appendix C: Digital Data Capture: Sampling

References

Abbreviations Used in the Guide

Bibliography

Interview Reports and Interview Instrument

Click here to download a printable pdf version.

Please send us your comments and feedback!

For HATII

Seamus Ross
Ian Anderson
Celia Duffy
Maria Economou
Ann Gow
Peter McKinney
Rebecca Sharp

For NINCH

President: Samuel Sachs II
President-Elect: Charles Henry
Executive Director: David L. Green

NINCH Working Group on Best Practices

Chair: David L. Green
Kathe Albrecht
Morgan Cundiff
LeeEllen Friedland*
Peter Hirtle
Lorna Hughes
Katherine Jones
Mark Kornbluh
Joan Lippincott
Michael Neuman
Richard Rinehart
Thornton Staples
Jennifer Trant**

* through June 2001
** through May 1999

Copyright 2002, National Initiative for a Networked Cultural Heritage

Version 1.0 of the First Edition, first published October, 2002

NINCH guide table of contents

Preface and Acknowledgements

I am delighted to introduce the First Edition of the *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. Since the Guide was first imagined and seriously discussed in 1998, much committed thought, imagination and expertise have gone into the project.

Back then it was clear that high-level guidance was needed (engaging multiple perspectives across different institution types and formats) to make sense of the plethora of materials coming out on information and technical standards, metadata, imaging, project management, digital asset management, sustainability, preservation strategies, and more. NINCH had been created in 1996 to be an advocate and leader across the cultural heritage community in making our material universally accessible via the new digital medium and this project seemed tailor-made for our new coalition.

Following NINCH's own good practice, the NINCH Board organized a working group to consider the best ways to proceed. That group is at the core of this project. We have lost and gained a few members along the way, but they are the *Guide's* heroes. Let me name them: Kathe Albrecht (American University), Morgan Cundiff (Library of Congress), LeeEllen Friedland (The MITRE Corporation, formerly Library of Congress), Peter Hirtle (Cornell University), Lorna Hughes (New York University), Katherine Jones (Harvard Divinity School), Mark Kornbluh (Michigan State University), Joan Lippincott (Coalition for Networked Information), Michael Neuman (Georgetown University), Richard Rinehart (Berkeley Art Museum/Pacific Film Archives, University of California, Berkeley), Thornton Staples (University of Virginia) and Jennifer Trant (AMICO). Archivists, librarians, scholars and teachers, digitization practitioners, visual resource experts, museum administrators, audio and moving-image engineers, information technologists, pioneers and entrepreneurs: all were represented in this group. Their expertise, good humor, persistence and good judgment have been essential to our producing this material.

After defining the project and declaring our core principles (detailed in the Introduction), the Working Group issued a Request for Proposals to conduct research into the state of current practice and to write the *Guide* in close collaboration with the Working Group. Of the several fine proposals submitted, we selected one from a broad and experienced team from the University of Glasgow. Under the leadership of Seamus Ross, a research team, based at Glasgow's Humanities Advanced Technology and Information Institute (HATII), mapped out an ambitious survey of the field for gathering information about current practice in the selection, planning, digitization, management and preservation of cultural heritage materials. We thank them for their work.

Although the Guide is the heart of this resource, you will see that we also include a general bibliography compiled by HATII together with the reports on the 36 interviews that formed the chief armature of the research underlying the Guide. I want to thank the 68 practitioners who offered us their experience and wisdom.

With a working draft in hand, the NINCH Working Group invited a team of volunteer, expert readers to consider our product. They probed and critiqued, and added richly to the text. Let me thank Melinda Baumann (University of Virginia Library), Stephen Chapman (Harvard University Library), Barbara Berger Eden (Cornell University Library), Georgia Harper (University of Texas), Sally Hubbard (Getty Research Institute), Leslie Johnston (University of Virginia Library), Amalyah Keshet (Jerusalem Museum, Israel), Deb Lenert, (Getty Research Institute), Kama Lord (Harvard Divinity School), Alan Newman (Art Institute of Chicago), Maria Pallante (Guggenheim Foundation) and Michael Shapiro (U.S. Patent and Trademark Office) for their readings and contributions. All who have read his comments would quickly agree with my singling out Steve Chapman as one who exceeded all of our expectations in the depth of his reading and the comprehensiveness of his responses. So a special thank you to you, Steve: we are indebted to you.

Julia Flanders, of Brown University's Women Writers Project, served as an inspiring copy editor, going far beyond what we might have asked of her.

Lorna Hughes, Assistant Director for Humanities Computing at New York University, arranged for the generous donation of web services to mount this edition of the *Guide to Good Practice* on the Internet. Antje Pfannkuchen and Nicola Monat-Jacobs have done a superb job of tirelessly mounting many pre-publication versions of the text online leading up to this final First Edition: we thank them heartily for their accurate and prompt work. Meg

Bellinger, Vice President, OCLC Digital & Preservation Resources, has offered the services of that division in mirroring the Guide on OCLC web sites in the U.S. and abroad and in furthering the Guide's further development. Thanks to Robert Harriman, Tom Clareson, Judy Cobb and Amy Lytle in making that happen.

Many thanks to the Getty Grant Program for initially funding this project and making it possible.

For all of its richness and complexity, we propose this as the first of several editions of a living document. Future developments and discoveries will add to and refine it. What can your experience add? The Second Edition will incorporate not only your comments but also an online navigational system based on a set of decision trees that should dramatically improve access to the information and advice.

Please use our Comments Form to update or correct information or suggest features that will enable us to make the Second Edition increasingly useful in assisting this broad community to network cultural resources more effectively.

David Green
October, 2002

NINCH guide table of contents

valid xhtml 1.1
abp~03/03

I. Introduction

The Case for Good Practice

Early developers of digital resources often had little thought for how their projects might dovetail with others. Today many of these projects suffer from this lack of forethought; they cannot be extended for broader use, they cannot be built upon by others and the chances are slim that they will survive into the future. More recently, the cultural community has begun to realize the importance of applying technical and information standards intelligently and consistently. The use of such standards not only adds longevity and scalability to the project's life cycle, but also enables an ever widening public to discover and use its digital resources.

One of the goals of this *Guide to Good Practice* is to show the critical importance for the community of moving beyond the narrow vision of these early project-based enthusiasts and thinking through what is needed to establish sustainable programs. By adopting community shared good practice, project designers can ensure the broadest use of their materials, today and in the future, by audiences they may not even have imagined and by future applications that will dynamically recombine 'digital objects' into new resources. They can ensure the quality, consistency and reliability of a project's digital resources and make them compatible with resources from other projects and domains, building on the work of others. Such projects can be produced economically and can be maintained and managed into the future with maximum benefit for all. In short, good practice can be measured by any one project's ability to maximize a resource's intended usefulness while minimizing the cost of its subsequent management and use.

By adopting community shared good practice, project designers can ensure the broadest use of their materials, today and in the future, by audiences they may not even have imagined and by future applications that will dynamically recombine "digital objects" into new resources.

Within the cultural and educational communities, there are today many different types of guides to good practice written for particular disciplines, institution types or specific standards. These include the Text Encoding Initiative's Guidelines for Electronic Text Encoding and Interchange, Cornell University Library's *Digital Imaging for Libraries and Archives*, the Digital Library Federation's *Guides to Quality in Visual Resource Imaging*, the Getty Trust's *Introduction to Vocabularies* and *Introduction to Metadata* and the UK's Arts and Humanities Data Service series of discipline-based "Guides to Good Practice." In creating the National Digital Library, the Library of Congress has been assiduous in providing documentation and discussion of its practices; similarly, the National Archives has published its internal "Guidelines for Digitizing Archival Materials for Electronic Access," and the Colorado Digitization Project has brought together in a web portal a wide-ranging collection of administrative, technical, copyright and funding resources.

Link Box:

Existing Good Practice Guides

Guidelines for Electronic Text Encoding and Interchange (Text Encoding Initiative): http://www.tei-c.org

Digital Imaging for Libraries and Archives (Cornell University Library): http://www.library.cornell.edu/preservation/dila.html

Guides to Quality in Visual Resource Imaging (Digital Library Federation): http://www.rlg.org/visguides/

Introduction to Vocabularies (The Getty Trust): http://www.getty.edu/research/institute/vocabulary/introvocabs/

Introduction to Metadata (The Getty Trust): http://www.getty.edu/research/institute/standards/intrometadata/

"Guides to Good Practice" (Arts and Humanities Data Service): http://ads.ahds.ac.uk/project/goodguides/g2gp.html

"Guidelines for Digitizing Archival Materials for Electronic Access" (National Archives): http://www.nara.gov/nara/vision/eap/digguide.pdf

Various documentation from the Colorado Digitization Project: http://coloradodigital.coalliance.org/toolbox.html

The Library of Congress has published many supportive materials; some notable resources include:

"Challenges to Building an Effective Digital Library": http://memory. loc.gov/ammem/dli2/html/cbedl.html,

"Technical Notes by Type of Material": http://memory.loc.gov/ammem/dli2/html/document.html

"Background Papers and Technical Information": http://memory.loc.gov/ammem/ftpfile.html

"Manuscript Digitization Demonstration Project, Final Report": http://memory.loc.gov/ammem/pictel/

"Lessons Learned: National Digital Library Competition": http://lcweb2.loc.gov/ammem/award/lessons/lessons.html

"Conservation Implications of Digitization Projects": http://memory.loc.gov/ammem/techdocs/conservation.html

Put simply, this plethora of information is daunting. Where does one start and how does one evaluate the relevance of any particular text in the growing corpus of material on project planning, digitization, the kinds of metadata that need to be included in any project, and the maintenance and preservation of digital resources?

As we detail below, the *NINCH Guide* has a good claim to being unique in providing a broad platform for reviewing these many individual statements. First, it is a community-wide document, created and directed by a NINCH Working Group culled from practitioners from digitization programs in different types of institutions (museums, libraries, archives, the arts and academic departments) dealing in different disciplines and different media. Second, it is based on a set of broad guiding principles for the creation, capture and management of networked cultural resources. And finally, it is also based on a set of intensive interviews of substantial digitization programs in the U.S. and abroad. The perspective is thus a new one.

By offering universal access to the knowledge this research brings together, the *Guide* should help to level the playing field, enabling newcomers to the field and projects which are smaller, either in terms of budget or scope, to offer resources that are as valid, practical and forward-thinking as projects that are created within information- and resource-rich institutions. It is this sharing of knowledge that truly facilitates the survival and success of digital resources.

History, Principles and Methodology of the *NINCH Guide*

The National Initiative for a Networked Cultural Heritage (NINCH) is a US-based coalition of some 100 organizations and institutions from across the cultural sector: museums, libraries, archives, scholarly societies, arts groups, IT support units and others. It was founded in 1996 to ensure strong and informed leadership from the cultural community in the evolution of the digital environment. Our task and goal, as a leadership and advocacy organization, is to build a framework within which these different elements can effectively collaborate to build a networked cultural heritage.

Realizing from the start the importance of connecting the big picture (the overall vision and goals for a networked cultural heritage) with actual practice within cultural institutions, NINCH board and staff concluded that organizing a comprehensive Guide to Good Practice was an important priority. A NINCH Best Practices Working Group was created in October 1998 to organize a review and evaluation of current practice and to develop a set of principles and guidelines for good practice in the digital representation and management of cultural heritage.

The Group proposed an initial definition of good practice by distilling six core principles from their own experience with a set of evaluative criteria to judge current practice. The Group thus proposed that Good

Practice will:

1. ***Optimize interoperability of materials***

   Digitization projects should enable the optimal interoperability between source materials from different repositories or digitization projects

2. ***Enable broadest use***

   Projects should enable multiple and diverse uses of material by multiple and diverse audiences.

3. ***Address the need for the preservation of original materials***

   Projects should incorporate procedures to address the preservation of original materials.

4. ***Indicate strategy for life-cycle management of digital resources***

   Projects should plan for the life-cycle management of digital resources, including the initial assessment of resources, selection of materials and digital rights management; the technical questions of digitizing all formats; and the long-term issues of sustainability, user assessment, digital asset management and preservation.

5. ***Investigate and declare intellectual property rights and ownership***

   Ownership and rights issues need to be investigated before digitization commences and findings should be reported to users.

6. ***Articulate intent and declare methodology***

   All relevant methods, perspectives and assumptions used by project staff should be clarified and made explicit.

With funding from the Getty Grant Program, NINCH issued a request for proposals to conduct a survey and write the *Guide*, in close collaboration with the Working Group. A team organized by the Humanities Advanced Technology and Information Institute (HATII) of The University of Glasgow was hired.

In order to ground the *Guide* in the reality of good practice that has been proven in the field, and to ensure that the personal views of the Working Group did not color the guide too much, the project began with a thorough review of current literature on the subject of good practice that included online and print resources, as well as gray[1] literature. This process was complemented by structured face-to-face and telephone interviews, and selective written exchanges with individuals from the cultural heritage sector.

The key information-gathering tool used for research was the *Digitization Data Collection Instrument for Site Visit Interviews* developed by HATII. For details on the development and use of this interview instrument see the "Introduction" to the Interview Reports. Interviews at digitization facilities lasted between 90 minutes and 3 hours and were conducted by four researchers on 20 site visits, involving 36 projects and 68 individuals from late 2000 through early 2001.

Sites were selected on a "best fit" basis to a matrix of project types and key themes established by the project team. The sites selected were not a scientific or representative sample, but as a group they broadly reflected the diversity of the community, while each represented one or more of the identified key themes of good practice. The rationale for site selection is further explained in the "Introduction" to the Interview Reports.

In parallel to the site visits, the research team undertook further focused research via literature review, telephone interviews and written correspondence on several broad themes: text encoding, digital preservation, asset management, rights management, and quality assurance. HATII identified another set of relevant digitization sites for inclusion in this stage of research. Theme reports written out of this research filled knowledge gaps that had not been addressed by the site visits and provided a more analytical view of current community good practice in these areas.

How To Use the *Guide*

The *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* is a unique contribution to the field. It takes a process-oriented approach to the digitization and management of cultural resources (keeping in mind their long-term life cycle from selection through preservation) and does so from a community-wide perspective. NINCH also intends to put into place a system for regular updates and further editions. The *Guide* takes the reader from the identification of available resources and the selection of material, through the creation of digital content, to its preservation and sustained access. For institutions that have not yet begun digitally representing material from their collections or making their born digital material accessible, the *Guide* will provide a way of coming up to speed in a quickly developing area. It identifies the decisions that need to be made, indicates when they need to be made and draws attention to the implications of the possible choices.

Users of the *Guide* will come from different backgrounds. Perhaps five examples will help you situate yourself among the possible categories of readers.

- If you are an archivist, librarian or museum professional, the *Guide* will help you select materials from your collections, reformat them, and make them visible and accessible to different audiences via the Internet or on portable digital media.
- If you are a funder, the *Guide* will give you an understanding of the activities involved in creating, delivering and sustaining digital content and background, and will help you to assess whether or not requests for funding are sensible and built on a thorough consideration of the issues.
- If you are an academic or other researcher, the *Guide* should give you sufficient information to design a project, convince collection owners to grant you access to material you need to digitize, and persuade funders to support your project.
- If you are a teacher of digitization in a library school or a faculty of information studies, the *Guide* can help you identify central issues to cover in digitization courses, and can provide your students with an understanding of the issues that they will need to address when they join a cultural heritage institution.
- If you are a vendor or manufacturer of software or hardware, the *Guide* should provide you with an indication of the challenges faced by the cultural community and of the significant levels of investment that the community is making in digital content creation, as well as showing you the tremendous value of the intellectual capital with which they are working.

This is not a recipe book for experts or specialists. It will provide content owners and decision-makers with sufficient guidance to know whether or not they are getting the best advice from their technical staff and whether their colleagues have put in place adequate strategies to guarantee the success of their digitization activities. It does not attempt to provide the final word on every topic, but instead supplies links to resources that we have evaluated and have concluded will offer a good next step.

The *Guide* identifies the decisions that need to be made, indicates when they need to be made and draws attention to the implications of choices made.

Humanities and cultural heritage institutions serve the needs of many different communities - from students and scholars to publishers and the general public. As you begin to develop and plan the use of digitization to make your collections visible and accessible, it is crucial to decide which audiences you aim to reach. This will influence many of your decisions: the items you select for digitization, the technologies you will use, and the mechanisms for delivering the digital materials to users. You may find, for example, that you have a collection that interests children as well as adults, but that each audience will require different delivery interfaces. While you could use the same technologies to reformat the material (and you would only need to do it once), and publish both versions using the same underlying delivery system, you would have to develop two separate interfaces to the same material.

Digitization may even change your sense of audience, by making it possible to offer broader access to rare or

inaccessible collections. Institutions often think first about digitizing material that is already popular with the public, but digital technologies now enable them to offer access to material that could not otherwise be seen or used, thus altering rather than simply reproducing the existing profile of use.

Audiences may be not only the users of the digital collections you produce, but also potential creators of digital surrogates from your collection for research, publication, advertising or enjoyment. Examples might be:

- an academic asking to digitize a collection of papers by a recently deceased contemporary artist as part of a research project
- a publisher proposing to produce a pay-per-view website with images of your collection of sixteenth-century engravings of native Americans
- a folk society requesting permission to include a rare recording of a 20th century storytelling from your collection on a CD they hope to release.

How do you respond to these requests?

- What best practices would you require if you were to agree to any or all of them?
- Would your expectations of each project be different or would you set them the same high standards?
- How would you ensure that, while you allow them each to use the material for their different purposes, you retain control of it in digital form, and that the processes involved in its digitization do not put the analog material at risk?

It is worth remembering that analog holdings constitute intellectual capital, and that as digital surrogates are created, the research, teaching or economic value of the originals should not be depleted. This may affect the material you choose to make accessible, the standard to which you do so, and what types of use and access arrangements you will put in place. Requiring those who work with your collections to follow good practices can minimize risks to the analog sources through their digitization.

So the first questions to ask include:

- Where is the audience for my collections?
- What types of individuals does that audience include?
- Will digitization enable me to meet the needs of existing communities better?
- Will digitization enable me to create new audiences for both the digital surrogates and the analog sources?
- What do I mean by “audience” in the digital world? Am I referring only to those individuals to whom I can deliver digital materials or am I also giving consideration to those who would like to produce digital surrogates for business, personal and research purposes?

---

[1] Gray literature, sometimes called "ephemeral literature," is unpublished material that can be lost to potential readers because it is not disseminated widely through publication or indexing. Examples of gray literature include: government or NGO research reports, workshop or conference papers, and theses.

II. Project Planning

Introduction

Planning is the first and arguably the most important step in any digitization project. Lest this sound like a platitude, it is worth noting that far too many projects are undertaken without adequate thought to the activities involved, the staff required, or the technical exigencies of the work. The need for good planning may be self-evident, but in practice it is often difficult to anticipate all the areas in which forethought is essential. Good planning for any project—even for managers who have successfully completed previous projects—requires a large number of decisions on questions such as the following:

- What work needs to be done;
- How it will be done (according to which standards, specifications, best practices);
- Who should do the work (and where);
- How long the work will take;
- How much it will cost, both to "resource" the infrastructure and to do the content conversion;
- Where, after having answered all of these questions, one might obtain funding.

This kind of planning is one of the most intellectually challenging of the project tasks, and may well be time-consuming. There may also be pressure to hurry this step, from a desire to show visible progress or in response to institutional pressure. But an investment in this kind of planning will be amply repaid over the life of the project: in the quality of the products, in smooth workflow, in staff morale, and not least in the total project cost. The goal of this section is to sketch out the parts of the planning process and indicate the important decisions—assessing the resources needed to complete the project, the staffing and equipment required, the choice and role of metadata, and the overall project management—and how to go about making them effectively. The checklist below gives a brief inventory of the resources required to undertake a digitization project. Not all projects will require all the resources listed, but this list will show the range of needs you should anticipate.

Technology develops and changes so quickly that decisions like those listed above may seem almost impossible to make with any confidence. Information on the array of standards, specifications, equipment, skills, and techniques not only presents a daunting learning curve, but also a welter of detail that can be very difficult to track. For the project planner, however, it is not these details that really inform good decision-making. It is much less important to know what sampling rate a particular piece of equipment offers than to understand how sampling works and how it can affect the quality of digital conversion. These underlying principles apply more broadly and change more slowly. Most importantly, though, they represent the level at which good planning takes place; with this knowledge, the planner has the tools to bring together an expert group of staff and consultants and create an effective framework within which they can work. This *Guide* contains detailed, up-to-date information on best practices in a number of technical areas, but the *Guide's* greatest and most enduring value for the project planner is its presentation of the more fundamental issues and how they interrelate.

The *Guide's* introductory section has already addressed the first question on the list above: What work needs to be done? By emphasizing the identification of audience and of your own institutional location and goals, the introduction contextualizes this decision and reminds us to ask "Who needs this work? Who will benefit?" The further ramifications of this question are explored in Section III on selecting materials, which discusses how to assess your collections and set priorities for digitization, and in Section XII on user evaluation, which provides guidance on how to assess the needs of your audience and how this information can shape your digitization strategy. This is also the stage at which you should get the facts and make your decisions concerning rights management, without which you cannot proceed with digitization: you need to establish the intellectual property status of the materials you wish to digitize, and you also need to decide on your own strategy for managing the intellectual property you are about to create. Both of these issues are explored in depth in Section IV. And although the project's final product may seem impossibly remote at this stage, you need to consider how the results will be distributed: not only what technologies you will use, but also how you will control access and ensure that you reach your intended audience. Section X covers these issues in detail.

The question of how the work will be done—the specifications, standards, and procedures you need to establish—has many facets which are addressed at various points in the *Guide*. Foremost among these is the question of standards: by using standards-based approaches wherever possible, you increase the longevity, portability, and interoperability of your data. You need to be aware of the standards that apply to the kinds of digitization you are undertaking, and these are described in detail in the sections on digitizing text, images, and audio-visual materials. Given the complexity and breadth of most standards, though, you also need to be aware of the best practices that apply to your community. For instance, both documentary historians and linguistic researchers use the XML-based Text Encoding Initiative Guidelines to encode textual data, but each group uses the standard in different ways that serve their particular needs. While you are considering the specifications for your data, you should also think carefully about how to capture and represent the metadata you will need to administer your digital materials and enable them to be used effectively. The *Guide* includes an appendix on metadata which describes the various types and their uses. The relevant sections of the *Guide* also provide pointers to specific information on best practices for particular digitization communities.

The question of "how" also involves decisions about equipment. For the project planner, these questions are most usefully addressed not at the level of specific brands and models, but by thinking about the functionality you require and the tradeoffs you are willing to make (for instance, whether keeping costs low is more important to the project's overall success than achieving the highest possible capture standard). The sections on images and audio-visual materials discuss how to approach these decisions; more specific information on particular kinds of equipment can be found in the appendix on equipment. Finally, you need to establish an effective workflow for your project. At the highest level, this includes project management strategies, which are discussed later in this section, and quality assurance methods (discussed in Section VIII). But in addition you need to consider how you will store, manage, and track your digital objects, which is addressed in detail in Section XIII on digital asset management.

Staffing issues—who should do the work—are closely related to the points just mentioned, since your decisions about methods and procedures may be difficult to separate from the staff resources you actually have available. Few projects have the luxury of hiring all new staff to an arbitrary standard of skill and experience. Further on in this section we discuss human resources: how to construct job descriptions and identify skilled staff, and how to set up a management and advisory framework that allows your staff the autonomy to do their jobs effectively. In Section IX, Working With Others, we consider a range of collaborative and cooperative relationships that may expand your staffing options, including project consultants, vendor outsourcing, collaboration with other institutions, and internal cooperation.

Once you have worked through the issues sketched above, you will be in a position to assess the practical scope of the project: how long the work will take, and how much it will cost. Of all the questions addressed here, these may be the most vulnerable to change over time, as techniques and equipment improve and grow cheaper, and as quality expectations rise. Some guidance on cost estimation is offered later in this section, and also in the sections on specific digitization areas (Sections V, VI, and VII). You should make sure in researching costs to take into account all of the startup and infrastructural costs the project will incur-costs for initial planning, choosing data specifications, building or choosing tracking and documentation systems, training staff, and so forth-as well as the incremental cost of digitizing the materials themselves. This is also an opportunity to consider the scope of your investment and whether this infrastructure can be reused or adapted for further digitization projects once this project is completed.

Finding the funds to undertake the project is the final step, at least logically; a successful funding request will almost always require a thorough consideration of the issues just described. Even if you are fortunate enough to have funding already committed, going through this process will ensure that you spend your resources prudently and receive value for your investment. Funding sources and strategies are discussed later in this section, and also in Section XI on sustainability.

The checklist box below gives a condensed list of the resources you may need to undertake a digitization project. Although not all projects will need all of the resources listed, it gives a sense of the range and options.

Checklist Box:

Resources that you will need for a digitization project:

Personnel:	advisors project management staff rights specialists researchers editors authors digitizers catalogers technical support/development legal advisors	☐
Software:	operating systems applications: --> image manipulation --> metadata authoring --> database --> indexing and search engine --> web server utilities server systems network clients specialist applications/developments	☐
Storage devices:	local hard drives network storage servers optical devices (e.g. CD writers) magnetic devices (e.g. tape drives) controlled storage environment	☐
Network infrastructure:	cables routers switches network cards ports	☐
Consumables:	stationery utilities printer cartridges lamps (for capture devices/special lighting) storage and backup media	☐
Project management:	preparing bids recruitment publicity and dissemination creation of deliverable product specifications design of workflow supervision of staff quality assurance	☐

Resources within your institution

If you are working within an institution that has other digitization projects under way, an examination of the resources already available within your institution is a good starting point. Staff will know if their department or unit has capture devices available or workers with experience of digitization or cataloging. This is an easy first step towards building a resource inventory, although knowing that you have one flatbed scanner, a digital camera and suitable equipment for digitizing audio, as well as people who know how to use that equipment, is

not on its own sufficient. A thorough identification of internal resources involves checking that:

- equipment and software are of a sufficient specification to meet your requirements;
- workers who can operate the equipment are available and appropriately trained;
- technical support and maintenance are in place;
- capture devices are (or can be) directly connected to your storage area; and,
- access to equipment and staff suits your workflow requirements.

Clearly assessing the adequacy of these resources is predicated on other decisions, such as your workflow requirements; indeed, many of the planning areas discussed in this section are closely interdependent. It should also be apparent why the *Guide's* introductory section stressed early on that you need to define what you want to do and the audience or audiences you intend to reach (see Section I). A clear statement of objectives (preferably in a formal document that can be shared with staff), combined with the resource inventory, will enable you to assess the suitability of your local resources.

You will make this document an even more effective planning tool by adding information about equipment specification (e.g. computer processor speed, RAM, hard disk capacity) and the results of tests for suitability. Before you can conclude that you have suitable resources you must test them to make certain that they will meet the requirements of the project. The Example Box below, "Resource Inventory and Test", shows what a resource inventory and test for scanners might look like.

Example Box:

Resource Inventory and Test:

PCs and Scanners	Functional Requirements	Suitability Test Result
1 Pentium 3, 600 Mhz, 128 MB Ram	Must handle processing and manipulation of image files up to 50 MB	Needs more RAM
1 Pentium 4, 1 Ghz, 384 MB Ram		***Okay***
1 Agfa Arcus		***Okay***
1 Agfa DuoScan 1200		Transparency tray inadequate

Overall Conclusion: Upgrade one PC and replace one scanner

Most large institutions in the cultural heritage sector will have resources that may be useful to the project but would not necessarily need to be borrowed for the entire life of the project. There may be physical equipment, such as devices for digital capture, analog capture equipment (e.g. record, tape, CD and video players that can be used when converting from analog to digital), network storage devices, or handling equipment and controlled storage for analog material.

Human resources may be even more useful—expertise in digitization, text encoding, networks or web delivery can often be found in-house. Even those institutions yet to carry out any significant digitization will have cognate areas of expertise. These skilled individuals can be difficult to find, so tell your colleagues that you are planning a digitization project and have them consider which skills might be of value to you. For example, the skills, techniques and processes required by digital photography are identical in many areas to analog photography, and the same applies to image processing. Similarly, the standards and methods for creating metadata have their roots in the creation of bibliographic records, library catalogs or finding aids and museum collection

management systems. In addition to this, it is important to consider the project team and project management process here. Projects should establish a set of procedures for project management from the very start of any project, identifying goals and time scales as well as tasks and outcomes tied to the availability of specific staff and equipment.

It is much easier to identify potential facilities and expertise within the framework of an institutional digitization policy or corporate technology plan—follow the more detailed questions for your own resources as described above. If such a policy has not already been adopted, it will probably be beyond the scope of an individual project to initiate one. Nevertheless, informal inquiries can still be made relatively easily. Remember that apparently unrelated departments or projects may be useful. For example, a great deal of high-end digital imaging takes place in dental, medical, biological and life science departments. The Internal Resource Identification Question Box illustrates some of the common areas of expertise to be found within an institution.

- Is there an institutional digitization policy to adhere to?
- Who else in the institution has digitization projects underway?
- What experience can you use (e.g., photographic, equipment analysis, etc.)?

Question Box:

Internal Resource Identification:

	Institution Type		
Resource	Academic	Library	Museum/Gallery
Imaging	Medical Imaging / Media Services / Photographic Services / Library	Special Collections / Photographic Dept	Imaging / Publications Dept
Metadata	Library	Cataloging Finding Aids	Collection Management Finding Aids
Text Encoding	Literature / Language / Computing Science Depts. / Information Management / Library	Cataloging / Information Management Finding Aids Electronic Texts	Finding Aids / Information Management

External Resources

Identifying resources outside your immediate department, unit or institution can be a more difficult process. Success depends upon what type of institution you are, your strengths and limitations, the accessibility of the resources you are seeking, and whether there is scope for collaboration. Guidance from and access to the experience of others are likely to be readily available. The Link Box points you to national organizations that provide information to support digitization projects. Outsourcing can be another way to fill gaps in the resources available locally, by contracting with a vendor, hiring a consultant, or establishing a cooperative relationship with another institution. These options are discussed in greater detail in Section IX, Working with Others.

Link Box:

Links to National Organizations Offering Guidance

CLIR: Council on Library and Information Resources: "The projects and activities of CLIR are aimed at ensuring that information resources needed by scholars, students, and the general public are available for future generations." http://www.clir.org/

DLIB Forum: "The D-Lib Forum supports the community of researchers and developers working to create and apply the technologies leading to the global digital library." http://www.dlib.org/

LOC: Library of Congress: "The Library's mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations." http://www.loc.gov/

NINCH: National Initiative for a Network Cultural Heritage: "A coalition of arts, humanities and social science organizations created to assure leadership from the cultural community in the evolution of the digital environment." http://www.ninch.org/

RLG: Research Libraries Group: "The Research Libraries Group, Inc., is a not-for-profit membership corporation of universities, archives, historical societies, museums, and other institutions devoted to improving access to information that supports research and learning." http://www.rlg.org/rlg.html

PADI: "The National Library of Australia's Preserving Access to Digital Information initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access." http://www.nla.gov.au/padi/

AHDS: Arts and Humanities Data Service: "Create and preserve digital collections in all areas of the arts and humanities." http://ahds.ac.uk/

HEDS: Higher Education Digitization Service: "The Service provides advice, consultancy and a complete production service for digitization and digital library development." http://heds.herts.ac.uk/

TASI: Technical Advisory Service for Images: "Advise and support the academic community on the digital creation, storage and delivery of image-related information." http://www.tasi.ac.uk/

Resource challenges

There are a number of challenges both in assessing and securing the resources required for the project. Projects that take place in large institutions frequently benefit from a significant amount of non-project-related investment. Such hidden benefits include local area networks, high bandwidth Internet connections, large capacity network-based storage devices, web servers, and technical expertise associated with maintaining and developing these facilities. This infrastructure provides the framework for the specific resources and skills a project needs, and without it many projects simply would never get off the ground. Although institutions are now trying to quantify this input, its actual value is difficult to establish, with the result that projects in well-resourced institutions are able to scale up more quickly but often under-represent the real costs that lie behind the their activities.

Equally, less well-resourced institutions and initiatives face an increasing challenge in matching the developments in presentation and delivery of digital resources that larger projects can provide. Frequently, the solution is for small and medium size institutions to develop collaborative projects. The Colorado Digitization Project (http://coloradodigital.coalliance.org/) provides a flagship example of how equipment, staff and expertise can be shared between large and small projects alike, enabling the digitization and delivery of resources that would not otherwise be possible.

Another challenge for digitization projects, large and small, lies in the area of human resources. Content creation is a burgeoning field and although many Internet businesses have failed, those companies such as Getty Images, Corbis, The Wall Street Journal and Reed Elsevier, which have adopted prudent content creation and marketing strategies, are showing steady growth. The finance, commerce, media and entertainment industries all recognize the value and benefits of digital assets, and this places a premium on skilled personnel. Furthermore, the development of staff with digitization skills related specifically to the humanities and cultural field has not kept pace with the growth in the number of digitization projects. Many projects report difficulties in recruiting and retaining staff. Few public sector projects can match the remuneration levels offered by the private sector, but there are strategies you can adopt that enhance your chances of meeting the human resources

challenge. These are outlined in the Human Resources Question Box.

Question Box:

Human Resources:

- Are there non-monetary factors that can be emphasized or enhanced? For example, will the project offer advantageous working conditions, training opportunities, or the possibility of gaining qualifications or accreditations?
- Are there aspects of the job that are more attractive than private sector equivalents (e.g. greater creativity, responsibility, freedom)?
- Can posts be combined or split to make most effective use of existing skills?
- Can you consider applicants from a non-humanities/cultural background, particularly for technical posts?
- Can any staff be re-deployed, temporarily transferred or re-trained from elsewhere in your institution?
- Can posts be shared or joint funded with other projects?
- Are you able to outsource any jobs?

Funding

Some project staff will be preoccupied with securing adequate financial resources to start, develop and sustain a project throughout its lifecycle. An accurate picture of the financial costs will help you to identify the financial pressure points and to estimate more accurately the overall costs of running the project. The sections below on skills, equipment, and project management will provide points to help you develop accurate project budgets. An accurate profile of project costs helps to minimize the financial unpredictability of the project and improves the probability that it will attract funding. Funding agencies remain attracted by the opportunities for funding initiatives in the heritage sector. The Link Box provides pointers to some major US funders.

Link Box:

Potential Funders of Digitization Projects:

- Andrew Mellon Foundation: The purpose of the Foundation is to "aid and promote such religious, charitable, scientific, literary, and educational purposes as may be in the furtherance of the public welfare or tend to promote the well-doing or well-being of mankind." http://www.mellon.org/awmf.html
- NEH: National Endowment for the Humanities, "an independent grant-making agency of the United States government dedicated to supporting research, education, and public programs in the humanities." http://www.neh.gov/
- The Getty: "The Getty Grant Program provides support to institutions and individuals throughout the world for projects that promote the understanding of art and its history and the conservation of cultural heritage." http://www.getty.edu/grants/
- IMLS: Institute of Museum and Library Services, "an independent federal agency that fosters leadership, innovation, and a lifetime of learning." http://www.imls.gov/grants/index.htm
- NHPRC: National Historical Publications and Records Commission, "supports a wide range of activities to preserve, publish, and encourage the use of documentary sources relating to the history of the United States." http://www.nara.gov/nhprc/

From the projects surveyed it is evident that most potential funders, particularly in the public sector, require applicants to provide a robust and auditable cost model. How this should be presented may vary from one funder to another, but it can be extremely useful to break down equipment and salary costs on a per unit or work package basis. Not only does it help the potential funders to make comparisons of unit costs between

projects within and across heritage sectors, but it also forces you to look at the process and scheduling of work in detail. The accuracy of these figures will be greatly improved by conducting a pilot study or by adopting a cost model from a previous project, even if it needs to be revised in light of the experience of the earlier project.

All the projects surveyed obtained their financial backing from a combination of institutional budgets, public grants, private donation or corporate sponsorship. None of the projects reported serious under-funding, although some found that the distribution of funds created an uneven cash flow, resulting in medium term planning problems. Similarly, none of the projects reported serious concerns about sustainability, even where the source of future funds was unclear. The general absence of plans for self-generating funds or of exit strategies supports this confident view that income would continue to materialize in the future. A number of projects have recognized that failing to adopt long-term financial planning is less than prudent. We recommend that time and support for securing further external funds are crucial as well as exploring the potential for self-generating income. Projects should develop an exit strategy that will secure the maintenance and accessibility of the digital material. These issues are discussed in more detail in Section XI on Sustainability.

Cost models

Determining the cost of digital content creation on a per unit basis is extremely problematic. Not only are there no comprehensive cost models available that cover all resource types but trying to apply such a model to the variety of institution types, financial arrangements, prevailing market conditions, nature and volume of material and the resolutions required would be problematic. Furthermore, the cost basis for creating, storing and delivering digital resources can be quite different and trying to establish a single cost per unit can disguise these differences or ignore them altogether. In spite of these problems it is possible to establish some bases for per unit cost.

At the simplest level a project can take the total funding required and divide it by the total number of units that they intend to digitize. For example total project funding of $300,000 divided by 40,000 units equals $7.5 per unit. However, such a figure can be extremely misleading. Firstly, there will be variation in per unit cost according to the type of material digitized. The creation of OCR text pages will differ from reflective color still images, which will be different again from 16mm moving images or 78 rpm records. Even within material of the same broad type there will be variation. Black-and-white negatives are likely to be more expensive to scan than black-and-white prints, since tone reproduction needs to be set image-by-image in the former case, while the same settings can be applied to a group of photographic prints. Even if a project is dealing with material of a uniform medium and size, variations can occur that impact on unit costs. A collection of bound, legal-size books may have volumes that cannot be opened beyond a certain degree for conservation reasons. This may require a different capture technique, for example capturing pages from above rather than inverted. Some volumes may have details that demand a higher capture resolution than the rest of the collection, while others may require curatorial intervention to prepare them for digitization. The extent to which projects need to take account of such details will vary but at the very least different material types should be distinguished as well as same-type materials that require different capture techniques.

The cost items that go to make up a per unit calculation also require consideration. Should pre-digitization conservation work, handling time, programmers and management staff be included in addition to capture equipment and staff? In practice, projects need to do both. This is best achieved by calculating the costs directly related to capture on a per unit basis, which facilitates comparison and cost effectiveness for different techniques. Non-capture-related items could then be added to provide a total project cost and a second per unit calculation could be carried out if required. The list box below provides an indication of how these different factors can be differentiated. It is common practice to calculate costs for audio-visual material on a per minute basis.

List Box:

Capture Cost Factors:

(per unit for a single media type with uniform capture techniques and settings). It is important to note that the digitization capture costs are actually the least costly of the whole process.

- Handling time (from the shelf to point of capture and return) as a percentage of total salary costs on a

daily basis
- Pre-digitization conservation work (this should only be applied for those items that have required it)
- Capture time (from set-up to naming and saving) provided as a percentage of the capture operators total salary costs on a daily basis
- Cataloging/Metadata (required for digitization and/or created at capture stage) as a percentage of total salary costs
- Hardware cost per item
- Quality Assurance time as a percentage of salary cost
- Software cost per item (both hardware and software costs should be on the basis of the depreciation of equipment or projected replacement cost, rather than the total cost of hardware and software)
- Hardware maintenance
- Technical support time (proportion of total salary or contract cost related to capture)
- Project Management time (proportion of total salary related to capture)
- Training (directly related to capture)

Non-Capture Cost Factors:

- Additional Project Management salary
- Web Programmer's salary
- Educational Officer's salary (or other additional project staff)
- Cataloging/Metadata (post capture creation) % of total salary costs
- Additional technical support salary
- Additional hardware and software costs
- Consumables (including any storage media such as CDs or DATs)
- Travel and per diem
- Training (non-capture related)
- Storage costs (based on total maintained cost for the gigabytes required)

Some sites with detailed information on costing are listed below.

Key Sites with resources on costings:

- Research Libraries Group: http://www.rlg.org
- Library of Congress: http://www.loc.gov
- Online Computer Library Center: http://www.oclc.org/home/

Human Resources

A project's long-term success depends on the accurate assessment of the required human resources, and producing a map of available and unavailable skills is a valuable starting point. Institutions vary in their areas of expertise and different types of project require different skills. Nevertheless, from the projects that we surveyed it has proved possible to develop a basic template of the people and skills required in realizing a digitization project. The requirements can be scaled according to the size of the project envisaged.

Job descriptions, performance indicators, training

Comprehensive job descriptions are indispensable, regardless of the project or institution. While job descriptions are not always required by the host institution, employment law often demands them. Funders are increasingly expressing an interest in viewing job descriptions as part of the application process as this provides them with a richer overview of the project. It is worthwhile developing an outline of job descriptions before the project reaches the recruitment stage. This is useful to determine the delegation of work, how jobs interrelate, which posts can be tailored to existing skills and which can be identified for external recruitment or outsourcing. A

useful process for developing accurate job descriptions is to set out a list of all the tasks required for a post and then rank them from highest to lowest priority or into essential, desirable and non-essential categories. Next, compile a corresponding list linking these tasks to the skills required, including any particular knowledge or qualification. Alongside this, compose a description of the experience or background required for these skills. Finally, review the original tasks and their priority to ensure that a realistic and coherent job description is produced. A resource which has been developed by the Association for Computers and the Humanities is a database of jobs in this field—it may be consulted by projects for guidance in drafting job descriptions, and can also be used to publicize new jobs to a focused audience of candidates. See http://www.ach.org/jobs/ for more information.

Example Box:

Sample Job Description

Job title: Digital Library Research Assistant

The Digital Library Research Assistant will play an integral role in the university's digital library projects, the goal of which is to bring a wide range of source materials to as large an audience as possible. The DLRA has responsibility for overseeing initial scanning and data capture, creating and reviewing metadata, and performing quality assurance checks. With other project members, collaborates on project publications and research.

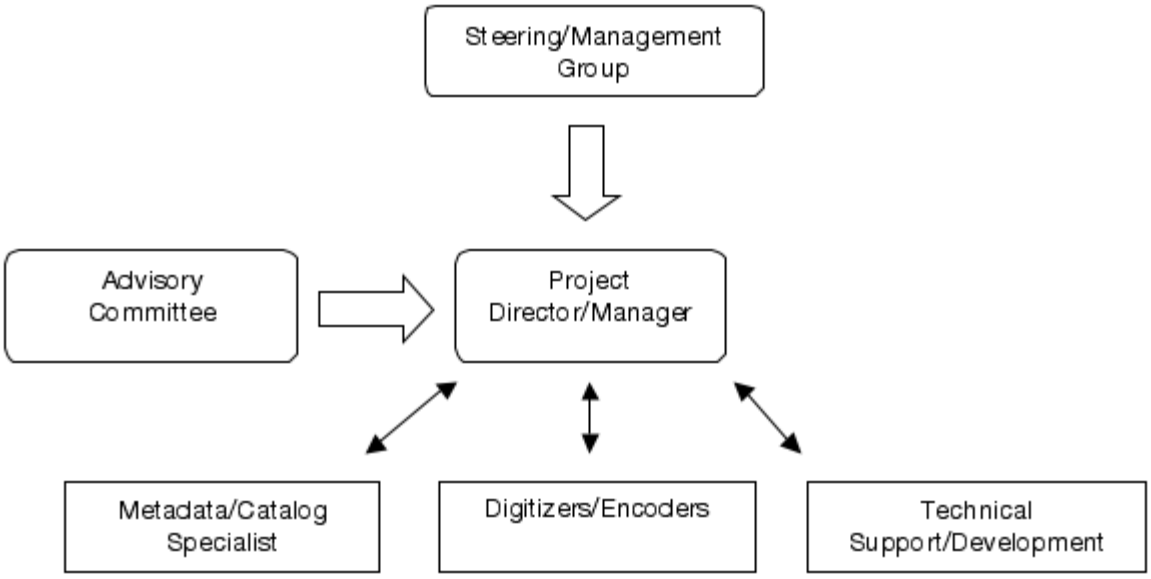
Job requirements: Bachelor's degree and one to three years' experience; basic computational skills, and expertise in at least one area of the humanities. Advanced degree and three to five years experience preferred. Familiarity with relevant encoding and metadata standards, including SGML/XML, METS and Dublin Core, is highly desirable. Must be a self-directed team worker with strong motivation and the ability to take initiative. Needs good communications skills (oral and written) and willingness to work collaboratively.

The use of performance indicators appears to be on the increase. They can have a positive impact, not least by providing a way of formally identifying training requirements. While most projects assess training needs on the job as an informal exercise, formal methods encourage appropriate training solutions to be planned and resourced in advance.

There is a close interplay between performance indicators, job descriptions and training assessments. The job description is very useful in developing meaningful performance indicators. Indeed, a useful starting point for performance review is to evaluate current tasks against those set out in the job description, highlighting whether the original job description was unrealistic, whether workloads need to be re-evaluated in the light of practical experience, or whether a skills shortfall needs to be addressed. The aim of addressing training requirements is to ensure that future tasks can be achieved and that the project will not encounter a skill shortage.

Managing the skills base

It is vital to ensure that a project be able draw on the right balance of skills. The challenge is to determine the skills of individuals and how they can most effectively contribute to the project. The key to successful delivery of projects is management. The diagram below incorporates elements from all of the projects surveyed, from the smallest to the largest, and illustrates the general structure that may be used to manage the project's skills base.



The steering group functions as an executive board and includes all constituents who are directly involved in the project, even if not employed by it, such as curators, archivists, subject specialists and education officers. In practice it is common for the steering group to be an existing committee within an institution.

The advisory committee is a broader-based group, providing general advice on the project's focus and direction. Members usually include the steering group with additional appointments from external organizations bringing particular areas of expertise, such as evaluation, to the initiative. There may be more than one advisory committee, or the advisory committee may be broken down into sub-committees each of which supplies more focused technical, academic or editorial decision-making support. This is the case with the Perseus Project at Tufts University, which has separate Technical and Academic Advisory Boards as well as a Steering Group to provide general project management. (Read Interview 28.2 for details on this arrangement)

It is essential to have a single project manager who is employed by the project, with responsibility for its daily management. In most cases the project manager provides the necessary project management experience, supplemented by internal or external advice. An institution needs to assign both accountability and authority to the project manager position, so that the process is not bogged down by myriad interactions with the advisory group or groups to deal with daily operations. In content creation projects it is unusual to employ external consultants to handle project management.

What skills are required?

There are four main areas, which will require staff with identifiable skills. These skill areas may be provided within a single project, dispersed across a collaborative project, or outsourced.

- Conservation: A crucial aspect of any digitization initiative will be a conservation assessment of the analog materials. Under some conditions this may show that before some material can be digitized it will require conservation intervention.
- Digitization/Encoding: This can involve digital imaging, keyboarding, OCR, character or full-text encoding, or a combination of these. In some projects it may also include conservation intervention in the analog material.
- Metadata/Cataloging: The creation of metadata records for the digital material. This work may also involve cataloging the analog material or searching for information to enhance the metadata record where it is absent from the analog version.
- Technical Development/Support: This falls into two distinct areas: the creation or implementation of specific IT solutions for creating, managing or delivering the digital material, and the provision of IT support for project hardware and software. This latter area includes desktop applications, network services, and capture devices.

In smaller projects staff may carry out tasks in more than one area: for example, the digitizer may also undertake technical development, or the project manager may take on metadata creation. In larger projects, such as SHOAH or the Genealogical Society of Utah, the duties of staff are so extensive that this is not feasible.

Project managers will have to decide whether to hire new staff with the required skills or to re-deploy existing staff from other areas of the institution. We found that many projects prefer the former, with two notable exceptions. First, there is a discernable trend for photographers to be employed for high-end digitization work. Projects have found that better-quality images are produced through training a photographer in digitization rather than trying to equip a digitizer with photographic skills. The second exception is the tendency to re-deploy or train existing cataloging staff in metadata creation. This is a logical progression for staff who will already have considerable experience in creating bibliographic records, collection management records, finding aids or catalogs, frequently in an electronic form such as MARC.

Another decision concerns background skills. With the exception of some technical posts, we noted a clear preference for staff with arts, humanities, library, museum or gallery backgrounds, or at least some experience or interest in the subject area of the collection. There may sometimes be advantages in not having such a specialization. For keyed-in text transcription, staff without subject knowledge are more likely to enter exactly what is on the page rather than interpret the contents and enter what they think is in text. On the other hand, subject knowledge can be exceptionally useful in gauging what areas of the content should be focused upon, deciphering difficult materials, or recognizing how areas of the content should be marked up.

When you are trying to find staff with appropriate skills, remember that some projects have benefited from using student labor and volunteers. The ability to draw on student labor represents a significant benefit for university-based projects. Projects such as those based at the University of Virginia Library have been able to build large and diverse digital collections because they are able to draw upon a pool of skilled, motivated and affordable labor. Projects that recruit student labor have invested considerably in training, adopted flexible working practices and tailored the work around the students' educational commitments. This approach has the added benefit of equipping students with the skill set required for future work, adding to the pool of available staff.

Volunteers often provide a similar pool of skills and projects such as the Genealogical Society of Utah have made effective use of this resource. They have found it both necessary and beneficial to invest in appropriate training for the volunteers. Such training should be factored into the project resource plans. In large-scale initiatives, volunteer management and training may become a significant part of the project itself.

The Link Box below provides links to sites that support skills development in digital representation.

Link Box:

An increasing number of organizations are offering training in digitization, which generally proves cheaper and far more useful than commercial training courses:

- Archive Builders: Document Imaging and Document Management. http://www.ArchiveBuilders.com
- Cornell University Library, Department of Preservation and Conservation: Moving Theory into Practice: Digital Imaging for Libraries and Archives. "This workshop is intended for librarians, archivists, curators, administrators, technologists, and others who are contemplating or implementing digital imaging programs." http://www.library.cornell.edu/preservation/workshop/
- HATII: Humanities Advanced Technology and Information Institute. Digitization Summer School: "The course will examine the advantages of developing digital collections of heritage materials, as well as investigate issues involved in creating, curating, and managing access to such collections." http://www.hatii.arts.gla.ac.uk/
- Humanities Computing Unit in Oxford: Summer Seminars covering a range of topics: http://www.hcu.ox.ac.uk/
- School for Scanning - North East Document Conservation Center: Creating, Managing and Preserving Digital Assets. http://www.nedcc.org
- Harvard Extension School, Museum Studies Certification Program. Muse E130: Information Technology for Museums. http://www.extension.harvard.edu/2001-02/courses/muse.shtml
- TASI: Technical Advisory Service on Images. Training courses "aimed at those involved in digitization projects, those who wish to capture images and those who wish to use digital images in teaching and research." http://www.tasi.ac.uk/training/training1.html

- UCLA/Getty course: Museum, Libraries and Archives: Summer Institute for Knowledge Sharing. http://skipper.gseis.ucla.edu/orgs/gettysi/html/summer.html
- University of Virginia Library: Digital Media Lab Services tutorials and short courses on digital image, video, and audio capture and editing. http://www.lib.virginia.edu/clemons/RMC/digilab-services.html

Equipment

Because our digitization capabilities are so strongly tied to—and limited by—the developing equipment technology, it is tempting to feel that the available technology should motivate our digitization strategies. However, on the contrary, it is vital to base equipment requirements on the characteristics of the collection/s to be digitized and on project needs, and not the other way around.

Although there are significant cost savings associated with outsourcing work to "offshore" production bureaus in Asia, the Far East, Mexico, etc, in cases where unique materials or special collections materials are to be digitized it is important that digitization should take place as close to the original as possible. Hence many projects will need to confront the complex questions of equipment specification and selection. A detailed discussion of matching material properties to hardware and capture settings can be found in Section VII on audio-visual materials. There is also further information on equipment choices in the appendix on equipment. At the moment we will focus on the basic differences in equipment and the technologies employed in order that the correct type of equipment resource can be procured for a project. Selecting the most appropriate equipment can be time consuming, but projects should not be deterred by the plethora of manufacturers and their competing claims. For example, the SCAN project (Scottish Archive Network) was initially unable to find a commercially available digital camera that exactly matched their requirements. Instead, they sourced a camera custom-made to their exact specification. This level of exactitude may be out of reach—and unnecessary—for most projects, but it is worth remembering that one need not be entirely constrained by what is commercially available.

Principles of digital data capture

Although there is a variety of capture devices for different applications, whether you are digitizing images, text, audio, video or 3D objects, the operating principles are the same. All digital capture devices take a sample of the analog source material to create a digital surrogate. This sample is made up of two elements: the sample rate and the sample depth. The sample rate describes how frequently readings are taken of the analog material. For example, in a digital image this would be the resolution, or the frequency per unit of area: the number of pixels per inch, expressed as pixels per inch (ppi) or dots per inch (dpi). An image captured at 600 ppi would have had 360,000 samples recorded per square inch. Similarly, for audio-visual materials the sample rate is the frequency per unit of time at which the source material is sampled. The sample depth is the amount of information recorded at each sampling point. For example, a sample depth of 24-bits would capture 8 bits for each of the three color channels (red, green and blue) at every sample point. For a more detailed explanation of sampling, see the appendix on digital data capture and Section VII on Audio-Visual Materials.

Selecting equipment

The medium, format, size, and fragility of the original material are among the primary factors affecting equipment choice. For text documents, flatbed scanners are suitable for single leaf, regular sized documents, provided the material does not go beyond the scanner's maximum imaging area (usually up to approximately US Letter size), or is put at risk by "sandwiching" it in the scanner. Large format flatbed scanners and sheet-feed scanners can handle single leaf, oversized documents. However, sheet-feed scanners put material at greater risk than flatbed scanners as the originals are pulled through a set of rollers. Drum scanners, whose imaging area is usually from 8" x 10" to 20" x 25", and digital cameras can also be used for oversize material, but they are an expensive option compared to flatbed scanners.

Bound pages that cannot be disbound, and pages in bindings that cannot open fully to 180 degrees require flatbed scanners with a right angle, prism, or overhead capture array. Digital cameras, with appropriate easels, book rests and weights are a versatile option for bound material. Camera beds or mounts, lighting, lenses, and filters all add to the cost and complication but make digital cameras more versatile tools for capturing manuscripts, bound volumes, original works of art, prints, out-size material and artifacts.

To achieve the highest quality scans of transparent media (e.g. 35mm slides and negatives, 6x4 and large format transparencies and microfilm) specialist equipment such as slide and film scanners, microfilm scanners or drum scanners should be used. Some flatbed scanners, with a dual light source, can handle transparent media though they often lack the dynamic range comparable to that supported by transparency scanners. However, you will not achieve as high a quality image as you would with a dedicated film or slide scanner. These have an inherently higher resolution, appropriate for the small size of the original, hold the transparencies more closely and securely, and frequently have negative color compensation to correct color casts for different types of film.

Audio and moving image materials present their own problems for digital capture. Not only is there a variety of source formats, including wax cylinders, 33, 45 and 78 rpm records, 8-track and cassette tapes, two-inch and VHS video in PAL and NTSC formats, but it is often very difficult to obtain access to analog devices for playback and linkage is difficult.

Definition Box:

Audio-Visual Facilities:

- Audio capture card required for sound material or video capture card required for moving images
- Source devices, such as 78 rpm record players and tape players.
- Mechanism for connecting these devices digitization equipment
- Intermediary device, such as a DAT (capable of handling ASEBU and SPDIF digital audio) machine

Link Box:

There are a number of audio and video digitization projects that are just getting started:

RAI: http://www.rai.it/portale

BRAVA: Broadcast Restoration of Archives through Video Analysis
http://www.ina.fr/recherche/projets/encours/brava/

COLLATE: Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material
http://www.collate.de/index.htm

PRESTO: Preservation Technology for European Broadcast Archives http://presto.joanneum.ac.at/index.asp

AMICITIA: Asset Management Integration of Cultural Heritage In The Interchange between Archives
http://www.amicitia-project.de/ami_home.html

The 3D representation of objects, from coins to buildings, is at the forefront of current digitization developments. At present the technology can be divided into two broad categories. The first, and simplest, is to create a moving image of an object. This is achieved by moving a digital camera around the object, or rotating the object in front of a fixed camera, while taking a series of still images. These images are then compiled to create a moving image of the object. The most common format for this is QuickTime VR. This is a reliable technology that requires a digital camera and mount or turntable. However, it does not provide a true 3D representation of the object because while only two planes are captured and displayed, it still represents 3D objects using two spatial planes. The viewer cannot manipulate the object, and the views provided are fixed and pre-determined.

Creating a true 3D representation of an object requires that the dimensions and features of the object be modeled. That is, the three dimensions of the object are represented in the computer as a set of coordinates. Attached to this "frame" are the textures of the object to provide the surface details. At present most 3D imaging technology remains in the sphere of industry. The technologies used to capture coordinates, render the model, and interact with the 3D representation (such as haptic feedback systems that allow one to "touch" the object, or 3D printing to create facsimiles) are often quite costly and require a relatively enormous amount of computing

processor power compared to the average desktop computer (in 2002). As such, 3D modeling devices remain application-specific, for example body imaging, prototyping or CAD/CAM applications. However, it was not long ago that digital imaging was the sole preserve of medical applications. During the next ten years we should see increasingly cost-effective and user-friendly devices that will bring 3D modeling into the mainstream.

Definition Box:

Virtual Reality:

Virtual reality can be described as an interactive, self directed, multi-sensory, computer generated experience which gives the user an illusion of participating in a three dimensional environment, even if a synthetic one. For cultural and heritage institutions, this may mean using virtual reality to create virtual representations of three dimensional objects in their collections or to create representations of environments, such as an Egyptian tomb, an ancient Persian palace, a historic Greek theatre or an ancient landscape. These three-dimensional objects could range from coins, vases, and sculptures to representations of whole rooms of collections.

Metadata

Metadata is an indispensable part of any responsible digitization program, and considerable attention has been paid to the definition of high-quality metadata standards for various purposes. (The appendix on metadata provides more detail on different types of metadata, and on specific metadata schemes and their uses.)The availability of accurate metadata is as important as the digital surrogates themselves for accessibility, usability and effective asset management. In many instances institutions will already have substantial metadata about the analog object (for instance, catalog records) much of which can be applied to the digital object. The project will be able to reduce its metadata creation costs by building on existing metadata. When selecting material for digitization you may wish to give priority to material for which partial metadata already exists.

It is crucial to remember to determine the status of the existing metadata, when you are assessing resource requirements. In an ideal world the existing catalog or finding aid would be complete and up to date. However, many libraries, archives and museums have a backlog of cataloging work, and part of a collection selected for digitization could fall into this category. Therefore, it may be necessary to devote time to locating missing information for your metadata records. You must then decide whether to seek information just for those fields required for the metadata, or to update the original catalog record in its entirety. Digitization provides an economical opportunity for institutions to expand their metadata, so consider the possibility of seeking extra funds or devoting more resources to this activity. Some of the new elements required for the metadata record of the digital object can be generated automatically: for instance, automatic metadata creation is a feature of much high-end digital camera software and of some OCR systems. Alternatively, a project may need to develop its own system, and can greatly improve the efficiency and accuracy of technical metadata. There is a general dearth of metadata tools, which poses a problem for the efficient creation and management of metadata for many projects. There is therefore likely to be a significant element of manual work, whether this lies in adding digital objects to existing electronic catalogs, creating records for web-based delivery such as Dublin Core, or implementing encoded metadata schemes such as EAD. Creating a metadata record will usually take as long as creating the digital surrogate and if detailed encoding schemes such as Encoded Archival Description or Text Encoding Initiative are used, this process can be considerably longer.

METADATA RESOURCES:

GENERAL METADATA RESOURCES

1. Canadian Heritage Information Network Standards Page: http://www.chin.gc.ca/English/Standards/metadata_intro.html
2. J. Paul Getty Trust, Introduction to Metadata: http://www.getty.edu/research/institute/standards/intrometadata/

3. Extensible Markup Language: http://www.w3.org/XML/
4. International Federation of Library Associations and institutions. Digital Libraries: Metadata Resources: http://www.ifla.org/II/metadata.htm
5. Text Encoding Initiative: http://www.tei-c.org
6. Metadata Encoding and Transmission Standard (METS): http://www.loc.gov/standards/mets/

METADATA MENTIONED ELSEWHERE IN THE GUIDE

1. Section III: Selecting Materials: Metadata & Interoperability. The Dublin Core metadata initiative http://dublincore.org/
2. Section IV: Rights Management: Technologies for Copyright Management and Protection.
    - The Open Digital Rights Language Initiative (ODRL): http://odrl.net/
    - Digital Object Identifier (DOI): http://www.doi.org
3. Section V: Digitization and Encoding of Text - Text markup schema. Text Encoding Initiative (TEI): http://www.tei-c.org
4. Section VI: Images
    - Descriptive:
        - Library of Congress Subject Headings (LCSH): http://lcweb.loc.gov/cds/lcsh.html#lcsh20
        - Categories for the Description of Works of Art (CDWA): http://www.getty.edu/research/institute/standards/cdwa/
        - Art and Architecture Thesaurus (AAT) http://www.getty.edu/research/tools/vocabulary/aat/about.html
        - VRA Core Categories: http://www.vraweb.org/vracore3.htm
        - Dublin Core Metadata Element Set: http://dublincore.org/documents/dces/
    - Structural:
        - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
        - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
    - Administrative:
        - A Web Hub for Developing Administrative Metadata for Electronic Resource Management http://www.library.cornell.edu/cts/elicensestudy/
        - Digital Library Federation, "Structural, technical, and administrative metadata standards. A discussion document:" http://www.diglib.org/standards/stamdframe.htm
5. Section VII: Audio and Video Capture and Management
    - Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
    - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
    - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
    - MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
    - Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
    - Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
    - Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
    - Library of Congress's Digital Audio-Visual Extensions to METS Standard http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
    - Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html
6. Section VIII: Quality Control and Assurance: Importance of Quality Control and Assurance of Metadata
7. Section X: Distribution:
   Metadata Harvesting
    - Clifford Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/mhp.html
    - Donald Waters, "The Metadata Harvesting Initiative of the Mellon Foundation," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/waters.html
    - The OAI MHP protocol: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
    - MHP tutorial: http://library.cern.ch/HEPLW/4/papers/3/
    - CIMI Working Group: http://www.cimi.org/wg/metadata/
    - CLIR Metadata harvesting project: http://www.clir.org/activities/details/metadata-docs.html
    - DLF and Metadata Harvesting: http://www.diglib.org/architectures/mdharvest.htm
    - University of Illinois at Urbana-Champaign Metadata Harvesting services: http://oai.grainger.uiuc.edu/

8. Section XIII: Digital Asset Management: "Metadata definition and management"
9. Section XIV: Preservation:
   Institutional Approaches
   - METS (Metadata Encoding and Transmission Standard) http://www.loc.gov/standards/mets/
   - UK Cedars project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
   - Australian Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metadata: http://www.nla.gov.au/padi/topics/32.html
   - NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
   - OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
   - Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf

Project Management

Many different approaches to managing projects are possible. While we found little evidence of the conscious adoption of a project management model, such as PRINCE 2 (http://www.kay-uk.com/prince/princepm.htm), most projects implemented many of the key features of successful project management. As understanding of digitization becomes more commonplace it may not be necessary to "hot house" prototype projects in the manner that many early projects experienced. However, it should also be recognized that integrating existing projects into host institutions often adds a layer of bureaucracy.

The Genealogical Society of Utah provides a good example of a comprehensive project management model. Each imaging project undertaken follows six stages:

1. Negotiation and project administration
2. Capture Convert Acquire
3. Image and metadata processing
4. Storage and preservation
5. Indexing and cataloging
6. Access and distribution

All projects will need to consider these six areas in setting up their own project management systems.

You do not necessarily need to adopt all the activities of a project management methodology; rather you need to scale the method to the needs of your project. The whole process should be determined by the project's objectives and rationale for creating the digital deliverable. Each process should be defined, together with the specific objectives to be achieved and activities to be carried out. The various roles and responsibilities should be detailed (defining job descriptions and breaking finances down aid in this — see above) and adapted to the size and complexity of the project. This should enable the efficient control of resources and facilitate regular progress monitoring. Regular reviews should be used to ensure that the project's objectives, which may change during the project lifecycle, are being met. Whatever project management method is adopted, it should provide a common framework and delineate milestones for all elements of the project.

In summary, your project management methodology should make possible:

- The use of pilot projects and feasibility studies to shape the overall scheme of activity
- Controlled and organized stages
- The establishment of a project plan with milestones
- Regular reviews of progress against plan and against objectives
- Control of any deviations from the plan
- The involvement of all constituents at the right time and place during the project
- Good communication channels between all constituents in the project and the sponsoring institution/s

Other key features are the need for one project manager to have ultimate responsibility and for the project advisory group to provide management quality control and assurance. In distributed projects, site managers are recommended in addition to an overall project manager. Most projects have relied on internal project management expertise, supplemented by external advice. Although many projects started as relatively autonomous there is a clear trend for project management structures and the project organization to be

integrated into the host institution's structure. This may be a natural progression for projects as they mature, but new projects may consider whether they should adopt it immediately.

Work flow and costings

While few of the projects interviewed carried out benchmarking tests most had conducted pilot studies. These were undertaken for a variety of reasons:

- Technical feasibility
- Technical forecasting
- Workflow analysis
- Training needs

When considering technical forecasting or prototyping, particularly in relation to costs, remember that there may be no corresponding benefit, and if there is a benefit it will vary for different types of content. Few projects in the humanities and cultural sector charge users for the digital deliverables. As such the cost/benefit may simply be realized by the ability of the project to amortize the depreciation on the equipment. A new high-resolution camera may pay dividends for fine textual or line art material, but not so for color images. Similarly, a device that enables the digitization of material that previously could not be captured, such as a 3D modeler, may not make financial sense if a project has to build in a profit or depreciation margin. However, if the device makes an important collection more widely available, the public access benefit may outweigh the financial costs.

Where any form of pilot study is undertaken it is important to build this into the project design and development cycle. For example, the University of Virginia Library's Special Collections department delineates its project work as intricately as possible before extrapolating its workflow and costings. This has given the project reliable data to forecast costs, but there are some areas where measurement has proved inaccurate, such as network transfer rates. The UVA Special Collections department also has a scheduling calendar tied to a tracking database to generate quality control and assurance checks and back-ups. In this respect it is typical of the projects surveyed which all use flowcharts, spreadsheets or Gantt charts to plan and monitor their workflow and costs.

If you are considering using a cost model (see above), it is important to include all the relevant costs, not just the obvious items such as equipment and staff time. You will also need to decide on what basis to evaluate — for example, costs per unit to be digitized or costs per hour. The table below provides a checklist of the factors that should be built into a cost model.

Finally, one further area to be aware of as you develop your cost estimates is digital asset management. In digitizing an image collection, for instance, you may well be generating a number of different kinds of digital objects-archival masters, delivery masters, thumbnails and other deliverables-which in turn will require storage, tracking, documentation, and upkeep. This process may require a significant commitment of resources and will need to be planned carefully. Section XIII covers digital asset management in detail.

Cost Model Factors	
Equipment	Purchase Maintenance Repair
Software	Purchase Upgrades
Staff	Salary (including benefits and insurance,) Training Recruitment Travel & subsistence
	Heat Light

Utilities	Water Phone Postage
Building	Rates Maintenance Upgrading/expansion

Analysis Box:

Costs of Digitization Programs

There is little information available about costs and this is an area where future work is necessary. In a rare exception, Steven Puglia analyzed the costs of digitization programs, in particular from the Library of Congress Ameritech competition and the National Archives and Records Administration's Electronic Access Report. The costs discussed are mostly *projected* and *estimated* costs, a problem discussed in the conclusion, which suggests that further studies are necessary. After an initial discussion on general costs of projects—it appears that on average, a third of the costs incurred by projects is the digital conversion, slightly less than a third is metadata creation, and slightly more than a third is made up of administrative and quality assurance tasks—the emphasis turns towards long term maintenance costs. The author suggests that these are not often taken account of with the project costs.

Three types of maintenance of digital objects are considered, each with mounting costs in relation to the initial costs per image:

- The first projection is an example taken from the maintenance of NARA objects. This involves only keeping one copy of master files and offering thumbnails and access online. The cost of this basic maintenance is in the region of 14-38% of the initial imaging cost.
- The second projection comes from Cornell report on computer output microfilm. Applying this to NARA, the cost would be 55-60% of the initial cost per image.
- The third projection looks at outsourcing to a commercial firm. This would cost 275-796% of initial cost. But it must be remembered that the other two costs do not include major IT infrastructure costs and thus are false figures, whereas the private firm costing will include this.

In conclusion, it is suggested that digital imaging may not be the best approach for long-term retention of information. Institutions can only justify retention if the images are used. Analog retention is the best way of holding materials in the long-term. In addition, it would be instructive to use figures from final project costs and also examine costs per person and production per person.

For the full report see: Steven Puglia, 'The Costs of Digital Imaging Projects', *RLG DigiNews*, October 15 1999, Vol. 3 No. 5. http://www.rlg.org/preserv/diginews/diginews3-5.html

Conclusion

At the start of any project, project planning feels like a way to exert control, eliminate risk, and guarantee a successful outcome. Certainly without good planning, the likelihood of failure and inefficiency is much greater. But you can be a better project planner by recognizing that the goal is not to eliminate risk but to prepare for it—not to control every variable but to create a project framework within which your team's response to the unforeseen will be resourceful and effective. In the technology domain, change and unpredictability are facts of life, and often represent opportunities rather than disasters for a well-planned project. Your planning goal should be to create a flexible, adaptable system whose staff and procedures can accommodate change. Your aim as a project leader should be to distinguish between what is essential—the central project objectives, the strategic components that will ensure long-term viability—and what is merely instrumental detail.

valid xhtml 1.1

III. Selecting Materials: An Iterative Process

Introduction

In this section we look at the issues you need to examine when selecting material - whether selecting physical originals for digitization, or reviewing born digital materials for preservation or republication. We also show how you can ensure that this process takes into account the aims and characteristics of your organization, the profile and needs of your users, and the characteristics of your collections. Some central questions you'll need to consider include the following:

- What are the aims of the digitization project or program and how can these guide selection, in cases where limited resources prevent you from digitizing the whole collection?
- How do these issues apply to material created in digital form?
- Who will be using these resources?
- What are their needs?
- Do the collections you selected for digitization need special handling?
- What are their characteristics and how do these affect the digitization process?

These are only some of the questions you will need to address when embarking upon a digitization project.

Characteristics of original material

Strategic knowledge of institutional holdings

Collections are a vital component of the intellectual capital of cultural heritage institutions and in most cases, their *raison d'être*. Successful digitization programs start from a strategic knowledge of the institution's collections and their relation to the institutional mission. A strategic assessment of this sort is also a vital step for regional, national, and international initiatives and collaborative or larger programs. Unfortunately, this prior analysis of existing holdings was often omitted in many of the early digitization projects.

> Planning for digitization should start from a study of the analog sources, the physical materials themselves, rather than in response to available technology or other pressures.

The analysis should include an assessment of the physical objects, their condition, characteristics, and educational, cultural, historical, and aesthetic value (Ross 1999).

Planning for digitization and digital publication should start from a study of the analog sources, the physical materials themselves, or, in the case of born digital material, the digital collections, rather than in response to the nature of the available technology or other pressures. As we are now passing from the first stage of experimental digitization projects to sustainable long-term programs, the technology itself holds fewer mysteries and attractions, and problems of scale and long-term planning come to the fore. It is thus important to carry out an overall assessment of institutional holdings before deciding on digitization priorities. In this way, even if digitization starts at a smaller scale with potential for expansion, its development will be well planned and will complement the institution's strategy and objectives.

Where strategic knowledge of the institutional holdings does not already exist, this type of assessment will require resources if it is to be carried out in a systematic and thorough way. It needs to be planned ahead, have institutional support at all levels, and include all team players. This can be demanding in terms of time and staff; however, the required resources will be well spent, as this analysis will be a valuable tool in planning all the organization's activities. It will clarify the strengths of the collection and will place them in a local, regional, national and international context.

When planning an evaluation of this type, physical access to the original materials is a fundamental consideration. This type of assessment needs to include both primary and secondary material and to examine the condition and size of both. A systematic assessment of the institution's collections will complement the resource inventory, as well as the institutional digitization policy we mentioned in Section II, highlighting particular strengths and resources in answer to the questions 'Who are you?' and 'What do you have?'

Intellectual value

For cultural institutions that are in the process of examining their assets, it is important to first assess their intellectual value to the intended audience. The project team, which should include custodians of the material, should assess whether the analog materials have sufficient intrinsic value and quality to ensure interest in their digital surrogates and to sustain the levels of access made possible by digitization. The Question Box below contains examples of the kinds of questions you may wish to ask when considering whether material should be selected for digitization.

Question Box:

Questions to Ask Before Selecting Material for Digitization:

- Is the information content of the analog objects high?
- What is the intellectual significance of the analog materials?
- How do they compare in terms of importance, authority, uniqueness, and timeliness?
- Is the object you are considering for digitization a good example of the period under examination?
- Is it original?

- Are there alternatives in the collection more capable of illustrating particular points?
- Is it complete or are there any parts missing?
- Is it up to date?
- Is it accurate?
- Would it withstand the test of time or is it only of ephemeral interest today? (Is it, for example, an important testimony of the past, of another culture or our own, of an artistic movement, an author, or a scientific discovery?)

As is clear from these questions, there is an element of subjectivity when making these judgments and weighing the intellectual nature of the various collections. Our perceptions and evaluation depend on our perspective and are subject to change. It is therefore advisable to consult widely within your institution and peer group as well as with your users in order to reach a general consensus in your decisions. Establishing the needs of potential users (see Section XII, User Evaluation), and determining what collections in other institutions might be digitized (see Section IX, Working with Others), will further enhance this activity, and can also help inform the institutional collection policy overall.

Added value and functionality

While examining the current intellectual value of the original materials, it is also worth considering the advantages of digitization in this area. A good example is the digitization of the Anglo-Saxon Beowulf manuscript, which was carried out by the British Library in collaboration with Professor Kevin Kiernan from the University of Kentucky with the close collaboration of scholars, curators, conservators, photographers, and technical experts (Kiernan 1994; Prescott 1998). This is one of the greatest treasures of the Library and of paramount importance for scholars of Old English. The manuscript had been badly damaged in a fire in 1731 and in order to protect the brittle and smoke-stained pages, each one was mounted in a protective paper frame in the mid-nineteenth century. In order to have a retaining edge for the frame, the letters around the edge of the verso of each leaf were covered, obscuring hundreds of letters from view. The digitization of the manuscript was one of the Library's early worldwide digitization projects, conceived in 1992 and completed with the production of a CD-ROM in 1997. Using a high-end Kontron digital camera (manufactured originally for medical imaging), while lighting each page with fiber-optic lighting, it was possible to capture digital images of the pages where the hidden letters were revealed.

In this case, digitization offered tremendous added value to the original manuscript, bearing in mind that some of those hidden letters represent the only known record of some Old English words. The project subsequently expanded to include a collection of digital images of all the primary evidence of the Beowulf text held in different geographic locations, ranging from the Royal Library in Copenhagen to Houghton Library at Harvard University, thereby creating a new resource and tool that allows enhanced study of the manuscript.

Digitization can also enhance the intellectual value of your original collections by allowing new possibilities for education and access, for example:

- by providing good quality digital surrogates of art works that are dispersed in different institutions around the globe;
- by allowing the creation of personalized study paths through the material, or juxtapositions of the works according to themes or particular interests;
- by bringing the collections to new audiences and allowing new ways of exploring and enjoying them;
- by offering the ability to search and manipulate text in electronic form, allowing a variety of ways to interrogate the material;
- by offering access to different versions and editions, opening up a richer and more complex experience of the text than is possible in printed form;
- by creating powerful electronic indexes which allow quick searching of video or audio clips and their synthesis in new forms.

Physical properties of source material

While the intellectual significance and content of your materials are very important, their physical characteristics also influence selection for digitization since they directly affect the digital outcome. Therefore, the analysis of the physical characteristics of the collections is an important step that will define how to handle the material, while deciding the subsequent digitization process. The custodians of the materials should ideally consult with those responsible for the digitization program to decide on what information is relevant. We include here some suggestions that are intended as a starting point to guide you in this process of examining the material itself and recording information about it. You might want to devise your own categories or expand on those listed in the Checklist Box: Physical Properties of Source Material.

Question Box:

Physical Properties of Source Material

- Type and category of object
  (e.g. Is it a book, manuscript, photograph, sound recording, TV broadcast?)
- Production process
  (e.g. Printed document, handwritten text, engraving, woodcut, wax cylinder recording, recording in mono or stereo? Is it an original or a reproduction/intermediary?)
- Date
  (e.g. How old is it? Do you have information about when it was made? If not, can you find out or estimate?)
- Physical size and dimensions
  (e.g. Is it a book with pages of regular (letter) size? Is it uniform? What is its length in cm, inches, duration of an audio-visual tape in hours/minutes/seconds, number of reels of film width; depth; thickness; weight?)
- Media Type
  (e.g. Paper, leather, wood, magnetic videotape, vinyl record; combination of materials; gold leaf details in manuscripts?)
- Format
  (e.g. 78 rpm disc, wax cylinder, reel-to-reel analog tape recording, DAT (Digital Audio Tape), Betacam SP tape, NTSC or PAL format video recording?)
- Sensitivity to light
  (e.g. What kind of lighting levels can it be safely exposed to during digitization? For how long?)
- Color information
  (e.g. Does it contain color information? Does color convey important information in this case? Is it an important element for the understanding and appreciation of the object?)
- Tonal Range
  (e.g. Does it have a wide tonal range? Is this an important element for the understanding and appreciation of the object/recording?)
- Noise
  (e.g. Does the audio recording have audio hiss, clicks and pops? Are there background sounds or images, which were captured in the original sound or moving image recording, that are not related to the main material? Is it important to preserve these?)

- Characteristics of born digital material
  (e.g. File format, resolution or sampling rate, bit-depth or rate, compression method, file size?)
- Characteristics and structure of informational content
  (e.g. For printed documents: does it include both illustrations and plain text? For sound recordings: what is the duration of the songs recorded and how many are included in the tape? For video recordings: what is the duration of the film recorded?)
- Structure of the material
  (e.g. Is the material bound or mounted?)
- Condition of the material and conservation
  (e.g. What is its state of preservation? Has it been assessed by conservators? Should it be conserved? Does it require any special handling?)

Digitization aims

In addition to the steps already outlined (examining the intellectual value of the materials, how this could be enhanced by digitization, and taking into account the physical characteristics of the collection), guiding principles for the selection process are the aims of the digitization program itself. These will vary between institutions, but we include here some of the most common general aims for starting a digitization program and how they might affect the selection of the material. In order to prioritize the selection process, examine which collections would be good candidates for:

- improving access;
- assisting the preservation of the original by reducing its wear and tear;
- supporting research activities;
- meeting user needs;
- allowing the use of good quality existing metadata;
- complementing other digitization efforts, either within the institution or beyond, without creating any rights management problems.

Increased access

In the examples of the Beowulf manuscript and the digitized art works previously cited, it is obvious how digitization can significantly increase access to resources. Examining whether digitization would indeed substantially increase resource accessibility is another criterion that can guide you in the selection process. This might involve asking the following questions:

- Does your institution already provide access to parts of the collection in non-digital forms?
- Is this easy and adequate?
- Does it serve your user community sufficiently? (This is further explored later in this Section — see User Needs and Demands.)
- Are there limitations to access due to the condition of some of the original collections?
- Are some of the original materials dispersed geographically within your institution or beyond?

The answers to these questions will reveal good candidates for digitization or will at least assist in establishing priorities. For example, some of the easily accessible analog material might move further down the selection list in favor of materials that are in remote stacks or storage areas. On the other hand, you might still select easily accessible materials, if they are so important to your users that the current level of analog access that you provide is inadequate. Other candidates might be significant but under-utilized collections with high intellectual value and relevance to the interests of your users. In this case it is worth further examining the reasons behind their limited use. Could you do anything to remedy the situation using methods more affordable than digitization?

When trying to identify a user group that would benefit from access to digitized materials, most institutions start from information collected about the current use of analog materials. The quantity, information needs, characteristics and location of current users can be a useful starting point for estimating future use. However, this is not always accurate, as digitization can increase access to and use of material that was hitherto unknown or underused. Additionally, the potential of digital information and web access can be so powerful and difficult to predict, that the actual users of digital resources are not always the same as those anticipated, especially where institutions are sharing digital surrogates of their assets with a worldwide community.

Preservation

Digitization of analog materials is not a substitute for investment in their conservation and preservation, but can assist the preservation of the original. Heavily used materials can benefit from the creation of faithful digital copies to prevent the deterioration of the originals, as can materials that are fragile or at risk. In this case, you should assess whether the benefit of digitization is greater than the risk placed on the material during the process of digitization. Digitization should also be a priority for cases where the existing storage medium is no longer suitable, as, for example, with nitrate film.

Digital files are themselves vulnerable to obsolescence, chiefly as a result of changing file formats, but also through deterioration of the storage medium. There are a number of strategies for overcoming this problem: careful specification of settings for quality capture and regular quality assessment; consistent use of well-defined and detailed metadata; and use of widely recognized, standard formats. These are discussed in more detail in Section XIV on preservation.

Research

Support for research activities is often an important incentive for digitization, although academics need to be aware that the resulting digital resources may not remain accessible indefinitely (see Section XIV: Preservation). Digitizing high quality, unique and original material can improve access for a wide community of researchers at all levels. It can also enable interdisciplinary collaboration and study from a wider range of scholars than was previously possible. For example, the creation of the Perseus database with its associated tools and resources has encouraged the study of ancient Greek and Latin culture in new ways, combining the analysis of texts, material culture, art, architecture, history, and geography.

> Interface design and appropriate metadata ... [require] considerable intellectual effort, knowledge, and resources to bring intelligence and context to the individual digital files, regardless of their quality.

While access to digital surrogates will never supersede the need for researchers to use the original objects, digitization offers added functionality that can assist research in other ways. It has some important shortcomings: it takes away the immediacy of the original and its impression of size and color (which are still not

displayed accurately with the existing technology today), and it eliminates or obscures some of the original context. The program team that selects and prepares materials for digitization can counterbalance these effects by providing sufficient context and accompanying information to make the digital objects meaningful. For example, good interface design and appropriate metadata can ensure that digitized illustrations do not appear on the screen in isolation, without reference to the rest of the text or the other related volumes; or that early LP recordings will not be played without reference to the text and images on the original disk jacket. Considerable intellectual effort, knowledge, and resources are required to bring intelligence and context to the individual digital files, regardless of their quality, which are little more than an electronic likeness of the original object. The program team must address these issues: it must examine the type and level of context it wants to provide, which should reflect the available resources as well as the research needs the project aims to support.

We have only begun to explore the possibilities of research tools for manipulating, searching, and studying digital resources. The ability to carry out sophisticated searches and accurate retrieval through rich materials is assisting all areas of research. In the area of text retrieval and processing, for example, the availability of large electronic corpora has contributed significantly to fields such as lexicography, socio-linguistics and authorship attribution studies. Developments in automatic translation programs and multilingual thesauri will allow much wider use of hitherto unknown resources and will enable cross-lingual searching. Image retrieval presents greater challenges. At the moment, few effective content-based image retrieval (CBIR) products have yet reached the marketplace. Most CBIR tools retrieve images based on appearance and the numerical characteristics of color, texture, and shape, rather than intellectual content and image semantics (Lesk 1998), and even in these areas they produce poor quality search results. However, CBIR continues to be the focus of intensive academic and commercial research activity. Even with the weak experimental applications that are available, including IBM's QBIC, Virage's VIR Image Engine, and Excalibur's Visual RetrievalWare, these tools clearly hold considerable promise to improve researchers' access to digital resources. Although there is still no substitute for trained catalogers, librarians, and subject specialists tagging the images with keywords, this is an area where future developments might revolutionize the use of images.

Link Box:

CBIR Tools to Watch:

IBM's QBIC: http://wwwqbic.almaden.ibm.com/

VIR Image Engine: http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/node41.html

Excalibur's Visual RetrievalWare: http://www.excalib.com/

A wider and richer collection of materials is also becoming available for research through advancements in even more challenging areas: development of next generation scanners for digitizing three-dimensional objects; the use of speech recognition for the indexing and retrieval of digitized speech archives; and research into video analysis and indexing systems that can parse and analyze hours of video, identify events as they occur, extract embedded textual data, employ continuous-speech recognition to convert spoken words into text and convert all this information into a searchable database. Collaboration between institutions (discussed in Section IX) also has a vital role to play in answer to the common current complaint that resources are not sufficiently large and comprehensive.

The digital environment continues to suffer from a number of limitations and restrictions that have a subsequent impact on research. For example, a large percentage of the material digitized thus far by institutions in the cultural and educational sector, belongs in the public domain. The selection has been driven more by copyright restrictions and less by the material's intellectual significance and usefulness for research and education. The rights to moving image and recorded sound material can restrict access to some information sources. Although OCR accuracy has been continuously improving for the last few years, the poor results with non-Latin characters has meant that very few texts in non-Western languages are available online (Smith 1999). Keyboarding is an alternative in these situations — not only for non-Latin alphabets but also for manuscript materials where OCR is not an option in any case. Several projects (for instance, the Japanese Text Initiative at the University of Virginia Etext Center, http://etext.lib.virginia .edu/japanese/) are creating resources in this way and using Unicode to handle the character set encoding. However, the added cost of these methods necessarily limits their size and scope. These issues go beyond technological development and are associated with socioeconomic conditions and pressures; they will need to be addressed by the research community as they have serious implications for the availability of resources and the direction of scholarship.

User needs and demands

User requirements should also guide the digitization selection process. Although it is difficult to predict future usage accurately in the electronic environment, current usage patterns of your collections can provide valuable pointers. Knowing the profile of your users and the ways they access information can highlight heavily used areas, limitations, and possibilities for development. Consider the following: What is the size of your user group? How are they distributed geographically? Is there a need for decentralized access of resources, e.g. from different institutions or from home? Will your users need special tools and facilities for using the digital resources? Do you have several different groups of users who may need different facilities and levels of access? If you do not already have sufficient information about your users to answer these kinds of questions, you will need a strategy for collecting that information and assessing whether digital resources can answer existing and future user demands. Digitization can also be used to bring in new types of audiences and help to open up specialized resources to a wider public. Digital collections can support lifelong education and learning, and they can also promote social and cultural inclusivity by opening up access for socially disadvantaged groups. If providing access to new user groups is one of the aims of the digitization program, it is important to try to involve the targeted users early on in this process. Consultation with these groups can bring new perspectives to your collections, assist in the selection of materials and provide information about new ways of using it. For example, providing schoolchildren with access to digitized resources from museums and galleries can be facilitated by early involvement with targeted user groups and discussions with schoolteachers about the ways that these can be used in the classroom. Working with different communities and ethnic groups can result in a variety of different ways of looking at the same material and making associations and links that go beyond the Western European, white middle-class biases.

In order to provide digital resources that are relevant and useful for your target groups, you will need to address the need for further evaluation and research. Evaluation involves user needs analysis and assessment using some of the methods we discuss in Section XII on user evaluation. This can take place before a digitization program begins, or before developing an existing program further. Information about how digital collections are being used is very limited at present and includes very little qualitative, in-depth detail that goes beyond the standard web usage statistics. Although we know, for example, that the American Memory web pages are used by millions of users every year[1], we know very little about how these resources are being used by a variety of audiences, even though this was one of the few projects that included an early evaluation survey of its pilot in 1991-3 with various schools, colleges, universities, and special, public, and state libraries (American Memory Evaluation Team 1993).

The level of technology available to user groups for accessing the digital material is another important consideration when selecting materials and approaches for digitization. If users do not have sufficient bandwidth to download or view large still or moving image files, for example, this material cannot be successfully disseminated. On the other hand, the digital environment encourages new expectations from users that your organization might not always be ready to meet. You will need not only to decide which are the most important needs to meet, but also to develop a clear articulation of the priorities and strategies that led to these choices.

Intellectual property and rights (IPR) management

As soon as you have identified appropriate material for digitization, you should tackle the intellectual property issues, since securing permission for materials you do not own can be a lengthy and often a costly process. If the materials you want to digitize are in the public domain, or if you own the copyright and control the digitization rights, then you can probably proceed without hindrance. Bear in mind that these issues—even concerning materials in your own collection—can be complex, for instance in cases where a photographer has been employed to take photographs of objects in the collection and these photographs are themselves to be digitized, or where gifts or bequests are concerned.

If the materials you want to digitize are not in the public domain, and if you do not control the copyright, you will need to identify the copyright holder and seek permission to digitize and publish the materials in question. You may already know who the copyright holder is, and may already have permission to use the materials for certain purposes, but be sure that this permission covers digitization and digital publication. Bear in mind as well that you may be asked to pay a fee or royalties for digitization or publication rights. In some cases, your needs may be covered by fair use (see more about fair use in Section IV on Rights Management) in which case specific permission from the copyright holder is not required.

It is possible that you may not be able to identify the copyright holders, or that they may not respond to your inquiry. If permission cannot be obtained even through good faith efforts, you need to proceed—if at all—very cautiously and assess the risks carefully. Again, in such cases, fair use may justify your inclusion of the materials in question. Restricting access to the materials may also limit your risk. Section IV includes more information on risk management, as well as links to resources on IPR.

Be very cautious, however, when considering publishing material on an unrestricted website. If what you plan to digitize cannot be justified as fair use, and if you cannot secure permission for digitization and electronic distribution of digital copies, you need to assess very carefully the risks of proceeding.

If, after assessing copyright status and fair use considerations, your institution finds that permission is required, you need to make a careful assessment of the value and costs that are at stake. Securing permission can be an arduous, time-consuming, and costly process, often with an uncertain outcome in cases where the copyright holder is either unknown or unwilling to give permission. The costs and uncertainty are frequently reasons why institutions decide not to digitize a particular work, despite its desirability. As indicated above, this raises the concern that some cultural materials will remain unavailable simply because their copyright status is uncertain or heavily restricted. Similarly, these materials may be available in digital form but only through commercial products, if copyright holders require fees that are beyond the reach of non-profit digitization efforts. Some have argued that cultural institutions with rich collections should avoid the effort and cost of securing permission to digitize materials they do not own, focusing instead on materials which they do control. (Zorich 1999), While this approach makes practical sense, it means that the shape and scope of digital collections will tend to reproduce the boundaries of the individual collection, rather than bringing together all related materials in a given domain, regardless of location and ownership. Collaboration can play a very important role in overcoming this limitation, but requires careful negotiation and coordination between institutions (see Section IX, Working with Others).

Even in the cases where your institution has been granted copyright to the analog material, there are further issues to be considered. You should examine, first, whether the rights you have actually extend to the digitization and distribution of electronic copies, particularly in the cases of donations and bequests. Even when there are good reasons for using materials to which you do not own copyright, you should be careful about the permissions or licensing agreements you negotiate and ensure that you have taken proper legal advice. Museums and art institutions should also be careful to protect the rights of the artists whom they display. Furthermore, when commissioning photographic documentation of your collections, you should ensure that the photographer's contract takes into account the digital environment and that you have cleared the rights for digitization and distribution of digital copies.

When you own the copyright of the materials you are selecting for digitization, make sure you plan to manage your rights efficiently and beneficially. It is important to include information on copyright status in the relevant metadata, and it is useful for the institution to have a rights management policy that includes guidelines on constructing a copyright statement to be displayed with the digitized work. Apart from the legal reasons for attaching a copyright notice, outlined in Section IV on Rights Management, a clear statement can deter users from misusing the material and, it can be argued, creates an implied contract between the user and the content provider to work within the confines of the statement.

In the desire to protect intellectual property, institutions and copyright holders alike often resort to methods which either compromise user access, or challenge the current concepts of fair use. Although a lot of research and work has been invested worldwide in technologies for protecting intellectual property in the electronic environment (also outlined in Section IV), many institutions still protect their intellectual property by making only low resolution images available on the Internet. Since these are not of sufficient quality to be used in commercial professional printed products, they are less likely to be worth misappropriating. However, although they may be adequate for various online purposes, they cannot support the kinds of detailed research for which high-resolution images are so valuable. Similarly, many owners of copyright materials feel threatened by the possibility of uncontrolled distribution of digital information, and are challenging the concept of 'fair use' in the digital environment, where it is more difficult to control who has access to the material and for what purpose. Fair use lies at the heart of the work of cultural and educational institutions, and many are striving to maintain it in at least its current form. However, you should note that generally, fair use is a safer option within a secured environment (a password protected class website, for example).

An institution must give serious consideration to how much control of its digital assets it is willing to give up or pass to others and examine how the costs involved in licensing and clearing rights would affect the overall budget.

Cultural institutions are exploring other mechanisms for managing the distribution of their intellectual property. Although none of these mechanisms offers a single uniform solution for the diversity of the cultural sector, some useful models are emerging. Licensing is the most popular option at the moment, but many institutions choose either to administer their materials directly, or to arrange for this to be done through some other means, whether through an external agency or through a consortium of rights holders. (Zorich 1999). Some of these options have existed for a long time, but are relatively new to cultural heritage organizations. The increasing demand for digital copies of the institutions' cultural assets, and the much wider distribution networks that the electronic world has brought, require new strategies from the cultural sector and careful examination of the different options. An institution must give serious consideration to how much control of its digital assets it is willing to give up or pass to others and examine how the costs involved in licensing and clearing rights would affect the overall budget.

Relation with other digitization efforts

With the spread of digitization activity and the multitude of projects and programs around the world, it is important to ensure that your proposed digitization efforts will not duplicate other efforts elsewhere. When selecting material for digitization you should ask yourself if there are any similar or complementary projects in the same institution, country, or even further afield, to avoid a costly waste of effort and resources. By collaborating and coordinating digitization programs, cultural institutions can ensure that they build a critical mass of digital collections. The community needs authoritative, well-governed registries for all types of cultural heritage materials —texts, images, audio collections, etc.—and easy ways to use the registries to determine whether items under consideration have already been digitized. Although such a registry already exists for preservation quality microfilm, there is still no similar listing of digital resources, but a number of initiatives in this area indicate that the situation is likely to change in the future.

Link Box:

Register of Digitization Initiatives and Programs

With the recognition of the need to eliminate redundancy of effort, to maximize the available knowledge about digitization projects and resources, and to improve communication between digitization activities, a number of initiatives in the US are providing mechanisms for projects to register details of their activities and outputs. Among those that have particular value for the humanities are:

- ***NINCH*** International Database of Digital Humanities Projects http://www.ninch.org/programs/data/
- ***ARL*** (Association of Research Libraries, Washington DC) Digital Initiatives Database (in collaboration with the University of Illinois Chicago) http://www.arl.org/did/
- ***DLF*** (Digital Library Federation) Registry of Digital Reproductions of Paper-based books and serials http://www.diglib.org/collections/reg/reg.htm
- ***CETH*** (Rutgers University Center for Electronic Texts in the Humanities) Directory of Electronic Text Centers http://harvest.rutgers.edu/ceth/etext_directory/
- ***IMAGELIB*** (University of Arizona Library) The Clearinghouse of Image Databases. http://www.library.arizona.edu/images/image_projects.html

Until initiatives in this area expand and such a source of information is created, cultural heritage professionals need to use traditional professional and research channels to collect information about existing digitization efforts and future plans. Web searches and relevant portals, word of mouth from colleagues, special-interest professional email lists, related professional and academic conferences and journals (e.g. *DLib* or the *RLG DigiNews*) can provide a wealth of information and a good starting point to map activities in the area.

Link Box:

Web portals on humanities computing resources

- ***Humanist*** discussion group http://www.princeton.edu/~mccarty/humanist/
- ***Humbul Humanities Hub*** (UK gateway of humanities resources) http://www.humbul.ac.uk/
- ***Labyrinth*** (Resources for Medieval Studies) http://www.georgetown.edu/labyrinth/labyrinth-home.html
- ***Voice of the Shuttle***, (the portal on humanities resources maintained by Alan Liu, University of California, Santa Barbara) http://vos.ucsb.edu/
- ***BUBL:*** Internet resources covering all academic subject areas, based in the UK. http://bubl.ac.uk/link/
- ***Accessing our Humanities Collections:*** a subject guide for researchers hosted by the JISC. http://www.jisc.ac.uk/dner/programmes/nff/humanities_collections. html

When carefully thought out and planned, collaboration with other institutions (covered in greater depth in Section IX) can be part of an effective strategy for digitization, enabling the sharing of resources and expertise. In fact, digitization activities have often encouraged and led to collaboration with others, as cultural institutions found that this was a way that they could afford expensive specialized equipment, or take advantage of the expertise of computing science departments to create advanced tools for managing digital collections, or enhance their digital collection by collaborating with institutions with complementary analog sources. By taking care that the materials you select for digitization complement initiatives at a local, regional, national, or international level, you can increase the breadth and impact of your work, making a valuable contribution towards truly useful digital resources and tools that answer real needs and are widely used. To this end, we need national or even, in the case of the European Union, trans-national strategies for the digitization of cultural material, to coordinate activities and ensure that we invest in creating unique, complementary, interoperable and high quality digital assets (Ross & Economou 1998).

Metadata and interoperability

Another criterion for selection of material for digitization is the availability and quality of the related metadata. Digitization activities often reveal backlogs and documentation gaps or inconsistencies in the management of the analog collections and the related analog documentation (which is also a form of metadata, even if not in digital form). It is important that you have good knowledge of the state of documentation, cataloging, and metadata used for the analog materials in order to make informed decisions about selection and to establish realistic projections of project costs and timelines. You may want to give priority to materials that have extensive, high-quality documentation and analog metadata in order to carry out the digitization activities without further delays, adding a minimum amount of digital metadata. On the other hand, digitization can provide the impetus for tackling cataloging backlogs or problems that hinder access to parts of the collection. In any case, a good assessment of the situation is important and should be undertaken before the selection stage, in order to allow you to plan both the cataloging and the digitization process accurately.

It is important that you have good knowledge of the state of documentation, cataloging, and metadata used for the analog materials in order to make informed decisions about selection and to establish realistic projections of project costs and timelines.

Another issue that influences selection is the existence of metadata about the digital surrogate. By coordinating not only with existing metadata activities within your organization, but also with national and international initiatives, you can avoid duplicating other efforts, and will be able to take advantage of the extensive work that has been taking place in this area. Have any other departments or staff in your organization carried out digitization projects before? Did they devise any subject categories or a metadata scheme that you could use and adjust? You need to decide on the type and depth of information you will record about the digital surrogate and how this will relate to the information on the original source. This has staff and resource implications and needs to be taken into account in the program planning stage. Selecting a uniform collection with materials of the same format and type, for example, will require less investment in metadata planning and recording.

There are three main categories of metadata: descriptive, administrative, and structural. You need all three to manage your digital resources. They are discussed in greater detail in the appendix on metadata, but some discussion will be useful here. Descriptive metadata is the information you will need to record in order to identify the digital resource, its analog original if not born digital, or any analog or digital derivatives. Administrative metadata is essential to the management of the digital asset, and describes its creation, intellectual property status, provenance, and the like: for instance, detailed information about technical specifications and the

digitization process which can ease future migration paths and ensure consistent results over time. Structural metadata describes the structure of the digital object and the relationships between its components; this information is crucial to assist navigation and ensure that complex objects that belong to a larger collection are linked meaningfully together. The recently developed METS (Metadata Encoding and Transmission Standard) encoding scheme, maintained at the Library of Congress, provides a robust way to record all three types of metadata, and allows the open use of multiple metadata standards simultaneously.

These metadata categories are vital for the longevity and preservation of the digital material, and in applying them you should work not only towards local consistency and completeness, but should also taking into account the work being done on metadata creation at a national and international level. Furthermore, if your digitized collections are to be retrieved easily by search engines and used together with complementary collections, consistent application of widely accepted standards in metadata recording is essential. Although there is no general consensus on the most appropriate metadata scheme for all metadata categories, it is important to keep abreast of the intensive activities in this area as generally accepted models start to emerge (OCLC/RLG 2001). The Dublin Core metadata initiative (http://dublincore.org/), for example, has gained international recognition for its efforts on descriptive metadata to ensure resource discovery and accurate identification in the electronic environment. It also demonstrated the importance, as well as the difficulties, of building general cross-sectoral consensus on metadata issues, which will be necessary in order to create resources that are interoperable across different user groups, disciplines, and institutional types.

(More information on metadata is available throughout the *Guide*, e.g. in Section V on Digitizing Texts, Section VI on Images, Section VIII on Quality Assurance, Section XIV on Preservation, and in the appendix on metadata.)

Digitization issues

Having examined all the areas discussed so far—the analog objects, their intellectual and physical characteristics, rights management questions, the reasons why these objects would be digitized, and users' needs—you should now assess the issues that all these raise for digitization.

In this process you should determine what features would have to be retained in the digital surrogate. Some of the questions you might ask are listed in the Question Box below. You should also examine those features of the original objects that might cause problems during digitization, such as the objects' size, their physical material, and their state of preservation.

Question Box:

What Features of the Original Should Be Retained in the Digital Surrogate?

- Should the digital copy retain color information?
- Do the images, sound, or video files you will deliver need to be of high resolution?
- Should digitization preserve the background 'noise' in the recording where this results from a live performance?
- Should digitization preserve the background 'noise' in the analog recording where this is a function of the age of the media or the process of production?
- Is it necessary to produce machine-readable text?

Examples of original features that may require special accommodation:

- Oversize maps will require special hardware or software for scanning
- Bound materials will need to be disbound before scanning, if using a sheet-fed scanner.
- Fragile items might require conservation treatment prior to digitization.

What are the technical and resource implications for digitization in order to retain these features or address these problems so that the results are of adequate quality to meet the aims of the program?

- You will need to calculate the file sizes that will be created from scanning in color and at high resolution.
- You will need to consider whether altering or destroying the original material is an acceptable option.
- You will need to assess the time and staff required for disbinding the bound books, and the availability and cost of large format scanners.
- You will need to assess the time, staff, and tools required to create descriptive, structural and administrative metadata.

When selecting a digitization approach, you should always start from your sources and users and aim to match these with the most appropriate procedure. Thus for example, it may not necessarily be appropriate to scan at the highest possible resolution or use a certain piece of equipment simply because it is available. If the technology you need is not accessible or affordable, you may wish to explore the possibility of collaborating with another institution that could offer access to the equipment you need. Alternative strategies, such as digitization from intermediaries, may also be worth considering.

You also need to examine the long-term use and development of the collection and estimate how it will grow. This should already be part of the selection process at an early stage, as future development and scope may influence your selection priorities and digitization strategies. If you know from the start that you will be dealing with a very large collection and will need to accommodate subject retrieval, for instance, you can build in provision for robust keywording even though this may seem like overkill for your first small collections. Although you might be starting small in the beginning, you will have prepared the ground for future expansion and growth.

Cost-benefit analysis

Another important consideration to take into account is the relationship of projected costs to expected benefits. Institutions in the cultural and educational sectors traditionally operate with limited resources and try to satisfy many competing demands. Although cost-benefit analysis is not the only consideration, it is nevertheless an important one that cultural heritage institutions cannot afford to ignore. However, what we should not forget in the analysis is that benefits from digitization might be intangible, especially when we are dealing with programs in the educational and cultural sector.

As digitization is very resource-intensive, at both the creation and the maintenance stage, you should first examine whether there are lower cost alternatives to digitization for achieving your stated goals. Sometimes, for example, the publication of a traditional scholarly catalog might serve your specialized research community better, or traditional access might be sufficient for resources that are of mainly local interest to a small group.

Even if you do decide to digitize, there might be lower cost alternatives to the digitization approach you have selected. For example, is color scanning really necessary to serve your aims and users? This is the most expensive approach (creating the largest files), compared to grayscale (which records shades of gray using 8 bits per pixel, as explained in Section VI on Images) or bitonal scanning (which records information only in black or white using only one bit per pixel). Bitonal scanning, which creates the smallest file sizes and is the generally most affordable approach, might be sufficient for scanning text, although there are documented cases of unbound grayscale scanning with a flatbed or sheetfeed scanner being less expensive per page than bound bitonal scanning with overhead scanners or digital cameras. Similarly, in some cases, uncorrected OCR text might be sufficient for retrieval and indexing. Hand-correcting the output of OCR

software improves retrieval accuracy, but significantly increases the costs of text conversion. It may even be worth asking whether you need to perform OCR in the first place, or whether it might be sufficient to provide an index with links to the page images.

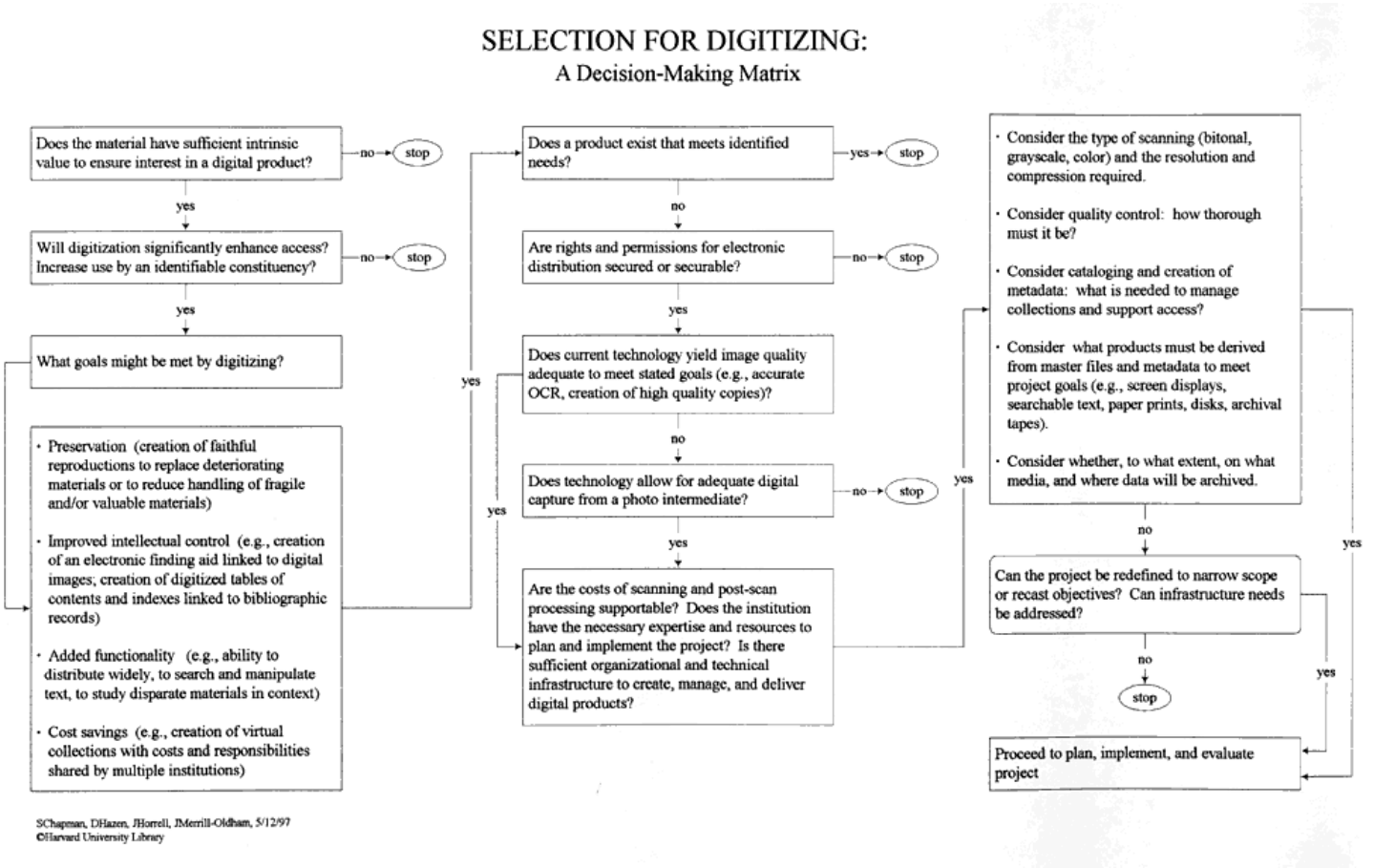
There are a number of challenges and even uncertainties in conducting a careful cost-benefit analysis. For instance, the labor costs associated with scanning are usually much lower than those related to selecting, preparing, inspecting, and indexing digital resources. Indeed, it is generally true that the intellectual overhead for digitization efforts—the cost of setting up infrastructure, making decisions, overseeing quality control, ensuring good quality metadata—is among the most significant costs. For this reason, it has been suggested that it is more economical to convert once at a high level in order to avoid the expense of duplicating the process at a later stage when more advanced technology requires better quality digital objects (Lesk 1990). However, these claims have not yet been borne out clearly in practice. Similarly, it can be extremely difficult to calculate expenses and compare the anticipated costs of new projects with those of existing projects. Again, the main costs are usually associated with the staff involved in digitization but this can vary enormously.

There also are many hidden costs that are often omitted in published reports. Cataloging, indexing, preparation of material for scanning, post-scanning processing of the material, quality assurance, and maintenance of digital resources are some of the activities that are not always calculated or indicated separately. “Though digitizing projects must calculate the likely costs and benefits, our ability to predict either of them is as yet rudimentary” (Hazen et al 1998). Another area of hidden digitization costs, ironically, is collaboration. Although, as we mentioned above, working with other institutions can potentially save money through the sharing of expensive equipment, it can also involve added commitments of time and resources—for instance, the overhead required in coordinating joint efforts, attending meetings, and the like. If not planned carefully, these may be missed in the project planning and create unforeseen cost overruns. Finally, in the calculation of costs and benefits other important factors are the levels of usage and the distribution mechanisms selected (the latter are discussed in greater depth in Section X on Distribution).

Digitization might lead to some cost savings, though these are unlikely to offset the project costs fully. For example, digitization may reduce the need for storage space, as is the case with the JSTOR project (http://www.jstor.org). This project helps the participating academic libraries reduce the costs associated with the storage and care of journal collections by digitizing and providing easier access to backfiles of over one hundred and fifty journals. Of course, this is no solution for museums where holdings are unique and cannot be discarded after digitization. Cost savings might also be involved in reducing staff time spent in retrieving information in analog form, but on the other hand, digitization creates a whole series of new demands and costs, particularly for the management and long-term preservation of the digital resources.

If digitization is indeed the best option, but involves higher costs than you can afford, you should examine whether you can secure external funding (see also Section XI on Sustainability). The priorities of funding bodies have often influenced the selection of material for digitization, with grant applicants trying to find out what is more likely to receive funding and selecting projects and material with that in mind. This is also a consideration with sponsors and donors of material, who often have their own agenda and can influence selection priorities. In some cases, however, you might find that the areas supported by external funding organizations agree with your institutional goals. Having a strategic knowledge of the collections and development plans, as discussed at the beginning of this section, can help you take a constructive approach in seeking sponsors and donors, so that their priorities match those of your institution.

The issues discussed here represent some of the questions you would need to examine with your project team when defining the selection criteria for your program. The criteria you identify will depend on the particular characteristics of your institution and the aims of the program. Once you have agreed on some key principles and guidelines, it is very useful to document them and share them among the team members. The diagram below, prepared by Harvard University Libraries, is one possible model which summarizes some of the questions and decisions involved in selecting material for digitization.



http://www.clir.org/pubs/reports/hazen/matrix.html or http://preserve.harvard.edu/bibliographies/matrix.pdf

SELECTION FOR DIGITIZING: A Decision-Making Matrix from Dan Hazen, Jeffrey Horrell, Jan Merrill-Oldham, *Selecting Research Collections for Digitization*, Council on Library and Information Resources (August 1998), http://www.clir.org/pubs/reports/hazen/pub74.html

[1] The American Memory Project received over 40 million requests per month in 1999, increasing to over 50 million in 2000, and over 80 million in 2001, numbers that exceed by far the number of readers who visit the reading rooms of the Library of Congress.

table of contents previous chapter next chapter

valid xhtml 1.1
abp~03/03

IV. Rights Management

Libraries want to share content; publishers want to sell it. Museums strive to preserve culture, and artists to create it. Musicians compose and perform, but must license and collect. Users want access, regardless of where or how content is held. What all of these stakeholders (and more) share is the need to identify content and its owner, to agree on the terms and conditions of its use and reuse, and to be able to share this information in reliable ways that make it easier to find. (Bearman et. al., 1999)

Introduction

Intellectual property rights (IPR) include copyright, trademarks, patents, publicity rights, privacy, and trade secrets, but it is copyright that will mostly concern this audience. Cultural institutions are primarily interested in two of the many issues that surround copyright: how they can legally digitize material in which they may not hold the copyright and how they can ensure that no one else can use the materials they have digitized without their approval (tacit or otherwise).

Copyright

Broadly speaking, copyright grants exclusive but limited rights to the creator of an original work to copy, reproduce, perform and distribute it. From its inception, however, copyright law has been as much about the promotion and circulation of knowledge and good ideas as it is about the protection and rewarding of creators. Limitations and exemptions to creators' copyright protection are as important to society as the protection itself. Certainly in the United States, the balance between the rights of creators and the rights of society in general is of key importance.

Categories of material

In the United States, "original works of authorship" protected by copyright include:

- literary works, which covers all kinds of literature (software is included in this category);
- dramatic works, including any accompanying music;
- musical works, including any accompanying words;
- artistic works;
- pantomimes and choreographic works;
- pictorial, graphic, and sculptural works;
- motion pictures and other audiovisual works, whatever their medium (film, television, DVD, etc.);
- sound recordings, which covers the spoken word;
- architectural works.

Practice and case law have tended to view these categories in their broadest sense. Maps are protected as ‘pictorial, graphic, and sculptural works’, and computer programs are protected as ‘literary works’. This categorization may affect certain rights; for instance, ideas, facts, short phrases, and blank forms are excluded from copyright protection, which focuses on the particular expression of ideas in a tangible medium.

Copyright and Digitization

Before carrying out any digitization, institutions need to establish the copyright status of the source material that will be digitized. If this investigation shows that the institution does not itself hold the copyright in the material, then the institution has three options: (a) abandon plans to digitize the material, (b) secure permission to digitize the material; or, (c) proceed with the project anyway under one of the exemptions to the exclusive rights of the copyright owner found in U.S. copyright law, such as the fair use exemption, but on the understanding that this involves an assessment of the risks of doing so (see Managing IPR Risk section, below). Projects using derivatives, such as photographic reproductions, rather than originals for digitization, need to examine the copyright status of both the derivative and the original. They may find, for instance, that by digitizing a photograph of a work of art they are infringing the rights of both the photographer and the artist who created the original work.

Once the project team establishes the rights status of the original and derivatives they plan to digitize, they should declare

this clearly in the metadata associated with the digital resource and/or the relevant web page.

When considering copyright ownership be particularly aware that there are layers of distinct rights in material in which a photographic reproduction, rather than an original, is being used to make a digital surrogate. The copyright status of both the original and the photographic derivative need to be determined. An excellent guide to negotiating the layers of rights related to photographs of visual materials is the “Copy Photography Computator” (http://vraweb.org/computator/welcome.html) produced by the Visual Resources Association.

Duration of protection and public domain

Any digitization project should begin with an analysis of who owns the copyright in the work to be digitized. If your evaluation of the copyright status of the material you hope to digitize reveals that your institution does not own the copyright, it may still be possible to digitize it if the rights in the material have passed into the public domain. For most categories of material, and particularly for literary and artistic works created after 1977, copyright protection generally lasts for ***seventy years*** after the death of the author/creator. When this term expires, the work enters the public domain where “all entities, information and creative works are available for use by anyone for any reason without restriction” (Zorich 2000). Note that there can be moral rights issues that exist independently of copyright term under certain foreign laws (see discussion of moral rights below). To avoid moral rights problems, avoid drastic cropping of visual works or other alterations and avoid removing the names of artists and authors of original work you reproduce.

The following chart prepared by Laura Gasaway, Director of the Law Library and Professor of Law at the University of North Carolina, summarizes the relevant terms and duration of protection according to the U.S copyright law, depending on the date when the material was produced or published. The chart makes clear the implications of the new copyright term that came into effect in the U.S. with the 1998 Copyright Term Extension Act. By extending U.S. term from 50 years after the death of the author to 70 years, and altering the special terms for certain published and unpublished works, this Act effectively put a twenty-year moratorium on works entering the public domain (Zorich 2000). For example, works published in 1923 that were next in line to enter into the public domain have been delayed until January 2019 as a result of the Act. Unpublished works (for example, historical documents like letters, diaries and manuscripts) created before 1978 will lose copyright protection in January 2003 unless they are published before that date. Previously unpublished works published between 1978 and 2003 are granted extended protection to December 2047.

WHEN WORKS PASS INTO THE PUBLIC DOMAIN

Includes material from U.S. Term Extension Act, PL 105-298

DATE OF WORK	***PROTECTED FROM***	***TERM***
Created 1-1-78 or after	When work is fixed in tangible medium of expression	Life + 70 years[1] (or if work of corporate authorship, the shorter of 95 years from publication, or 120 years from creation[2])
Published before 1923	In public domain	None
Published from 1923 - 63	When published with notice[3]	28 years + could be renewed for 47 years, now extended by 20 years for a total renewal of 67 years. If not so renewed, now in public domain
Published from 1964 - 77	When published with notice	28 years for first term; now automatic extension of 67 years for second term

Created before 1-1-78 but not published	1-1-78, the effective date of the 1976 Act which eliminated common law copyright	Life + 70 years or 12-31-2002, whichever is greater
Created before 1-1-78, but published between then and 12-31-2002	1-1-78, the effective date of the 1976 Act which eliminated common law copyright	Life + 70 years or 12-31-2047 whichever is greater

1 Term of joint works is measured by life of the longest-lived author.

2 Works for hire, anonymous and pseudonymous works also have this term. 17 U.S.C. § 302(c).

3 Under the 1909 Act, works published without notice went into the public domain upon publication. Works published without notice between 1-1-78 and 3-1-89, effective date of the Berne Convention Implementation Act, retained copyright only if, e.g., registration was made within five years. 17 U.S.C. § 405.

Notes courtesy of Professor Tom Field, Franklin Pierce Law Center

Chart prepared by Laura Gasaway, University of North Carolina,
Last updated 9-18-01, http://www.unc.edu/~unclng/public-d.htm

Work for hire

Although in some countries only an individual can be recognized as the author of a work, in some countries, such as in the U.S. and the UK, the author can also be an organization. Unless otherwise specified by contract, copyright in any material (whether literary, graphical or photographic) created by staff at universities, as part of their normal duties will belong automatically to the university as the employer. In addition to copyrights created by employees, employers can also own copyrights created by independent contractors when the work falls into one of the narrow categories of work defined by the copyright law as a “work for hire” and when there is a written agreement between the parties that the work will constitute a “work for hire.” It is likely, therefore, that if your institution were to commission a photographer to take a photograph of an original artwork, the copyright in the photograph would reside with the photographer unless specific actions are taken to transfer the copyright to the institution. For this reason, it is important to clarify these issues from the outset and include appropriate transfers of copyright in all relevant contracts. For works commissioned before digitization became a common practice, the current rights position must be considered, and in some cases re-negotiation may be necessary. If you are outsourcing digitization work, make sure that the contract specifies that you hold any rights resulting from the digital files created by the contractor.

Fair use

Cultural institutions may wish to digitize materials that are not in the public domain, whose copyright they do not own. In this case, they should examine whether the material and the way they plan to use it may be covered by ‘fair use’. ‘Fair use’ is an exemption under U.S. copyright law that allows one to legally use copyrighted material without the explicit permission of the copyright owner. Fair use is named differently by different national legislations. Different nations also vary on what can be covered by fair use, but these usually include non-profit educational use and private research and study. In the U.S., fair use is one of a set of exceptions to exclusive rights and is framed by four key factors:

1. the purpose and character of the use (e.g. whether such use is of a commercial nature or is for not-for-profit educational purposes; whether the use is transformative);
2. the nature of the copyrighted work (e.g. whether it is based on facts or is an imaginative work);
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole (a portion of a book rather than the whole, for example);
4. the effect of the use upon the potential market for the copyrighted work.[1]

In an effort to address the interpretation of these four factors, the Clinton Administration established the Conference on Fair Use (CONFU) in 1994. CONFU brought together intellectual property owners and users to negotiate new guidelines for the fair use of digital material in nonprofit, educational contexts. The Conference set up several working groups to investigate the issues, but when CONFU concluded in 1998, none of the groups had produced guidelines acceptable to all parties. Most major educational and cultural organizations were critical of the draft guidelines that had been prepared and opted not to

endorse them. In some cases, it was felt that the Guidelines asked the right questions, but provided the wrong answers.

Some institutions decided to review and adjust the Guidelines or design new ones of their own to meet their needs and fulfill a wider strategy of IPR management. One example is the Visual Resource Association's *Image Collection Guidelines: The Acquisition and Use of Images in Non-Profit Educational Visual Resources Collections*, a guide that was developed as a result of VRA's involvement in preparing the CONFU Digital Images Guidelines. Another example, adapted from the CONFU guidelines, are the short and clear "Rules of Thumb" for interpreting the fair use of copyrighted materials devised by Georgia Harper for the University of Texas System. Finally, a third example of an alternative to the CONFU Guidelines is the following "Checklist for Fair Use," prepared by the Indiana University Copyright Management Center (Indiana University-Purdue University Indianapolis).

Link Box:

Alternatives to the CONFU Guidelines

Visual Resource Association, Image Collection Guidelines: The Acquisition and Use of Images in Non-Profit Educational Visual Resources Collections: http://www.vraweb.org/copyright/guidelines.html

Georgia Harper, "Rules of Thumb", adapted from the CONFU Guidelines: http://www.utsystem.edu/ogc/IntellectualProperty/roftimag.htm

Indiana University Copyright Management Center, "Checklist for Fair Use": http://www.iupui.edu/~copyinfo/fuchecklist.html

---

Checklist for Fair Use

Please complete and retain a copy of this form
in connection with each possible "fair use" of a copyrighted work for your project.

Name: _____ Date: _____

Institution: _____ Project: _____

PURPOSE

Favoring Fair Use	*Opposing Fair Use*
☐ Teaching (including multiple copies for classroom use)	☐ Commercial activity
☐ Research	☐ Profiting from the use
☐ Scholarship	☐ Entertainment
☐ Nonprofit Educational institution	☐ Bad-faith behavior
☐ Criticism	☐ Denying credit to original author
☐ Comment	
☐ News reporting	
☐ Transformative or Productive use (changes the work for new utility)	
☐ Restricted access (to students or other appropriate group)	
☐ Parody	

NATURE

Favoring Fair Use	*Opposing Fair Use*

- ☐ Published work
- ☐ Factual or nonfiction based
- ☐ Important to favored educational objectives
- ☐ Unpublished work
- ☐ Highly creative work (art, music, novels, films, plays)
- ☐ Fiction

AMOUNT

Favoring fair use	*Opposing fair use*
☐ Small quantity	☐ Large portion or whole work used
☐ Portion used is not central or significant to entire work	☐ Portion used is central to work or "heart of the work"
☐ Amount is appropriate for favored educational purpose	

EFFECT

Favoring Fair Use	*Opposing Fair Use*
☐ User owns lawfully acquired or purchased copy of original work	☐ Could replace sale of copyrighted work
☐ One or few copies made	☐ Significantly impairs market or potential market for copyrighted work or derivative
☐ No significant effect on the market or potential market for copyrighted work	☐ Reasonably available licensing mechanism for use of the copyrighted work
☐ No similar product marketed by the copyright holder	☐ Affordable permission available for using work
☐ Lack of licensing mechanism	☐ Numerous copies made
	☐ You made it accessible on Web or in other public forum
	☐ Repeated or long term use

Prepared as a service of the
Copyright Management Center at Indiana University
located on the campus of IUPUI
http://www.iupui.edu/~copyinfo/
rev. 3.99

Created: 26 July 1999, ARK ***Last Updated***: 25 January 2001, LDB

Comments: copyinfo@iupui.edu ***URL***: http://www.iupui.edu/~copyinfo/fuchecklist.html

Copyright 1995-2001. The Trustees of Indiana University

Institutional guidelines can help project managers make consistent decisions on what can be considered ‘fair use’ in their everyday work. However, it is important to remember that it is ultimately a court of law that determines whether a particular use, when challenged, is ‘fair use’ or not, according to the particular circumstances of each case. Some uses which may seem as if they are clearly ‘fair use’ might be challenged.

Link Box:

For more information on 'Fair Use', good starting points are:

- Stanford University Libraries: Copyright & Fair Use page: http://fairuse.stanford.edu
- Duke University, Center for Instructional Technology: Copyright Law and Fair Use: http://cit.duke.edu/resource-guides/development-copyright.html
- The CONFU Final Report by Bruce A. Lehman, November 1998 is available on the U.S. Patent and Trademark Office website: http://www.uspto.gov/web/offices/dcom/olia/confu/

Moral rights

Even when a cultural institution or publisher owns the copyright in a piece of literature or art, in many countries the creator (e.g. the artist or author) retains the moral right (*droit moral*) to be identified as the creator of the work and is granted protection against derogatory treatment of his or her work. For cultural institutions this means, for example, that it is essential to ensure that the original artist's name be linked with his or her creation (and the use of appropriate metadata is very useful in this direction), that works be normally used in their entirety, and that they not be amended (e.g. digital copies should not be cropped or edited). These issues should be covered by institutional policy in clear guidelines for use. In the U.S., moral rights are limited to the 1990 Visual Artists Rights Act, which recognizes visual art authors' right of attribution, their right of integrity, and their right to prevent the destruction of copies of the work[2]. However, not all moral rights fall within copyright laws as in the U.S., and not all moral rights terms are finite. In some instances, as in France, moral rights may survive the expiration of copyright. Thus one must be careful not to alter or denigrate a reproduction beyond recognition or in a context that may be objectionable to the artist or artist's estate.

Seeking permission to use copyrighted material

In order to legally use material that is not in the public domain and is not covered by 'fair use' or otherwise exempt from liability for infringement, you must acquire permission from the rights holder(s) (unless the material has been licensed for your use and the permissions have been assigned as a part of the license). Depending on the nature of the work or the country of origin of the author or artist, there are numerous collective rights organizations that can help you to clear these rights. The Copyright Clearance Center (CCC) (http://www.copyright.com) for parts of books or journals is one such example.

Link Box:

The Getting Permission page by Georgia Harper at the Office of General Counsel, University of Texas System (http://www.utsystem.edu/OGC/IntellectualProperty/PERMISSN.HTM) includes links to collective rights organizations you can contact to gain permission about the following types of materials:

- Images
- Journalism
- Music Performances
- Plays
- TV News Programs
- Movies

It also includes links to resources for tracing the rights owners of archival materials, such as historical photographs, architectural drawings, or personal papers.

When you do not know who the rights owners are and colleagues cannot offer any additional information, you can contact the U.S. Copyright Office (http://www.loc.gov/copyright/), which provides online searching of some of its records on registration and copyright ownership documents. It also carries out professional searches for a fee. Once you know who the rights owner is, you can contact them directly to ask for permission to use the material. Apart from your contact details, you should describe clearly:

- the work you want to use (with a copy if possible);
- the scope of the project;
- where and how you intend to use the work (e.g. names of key contributors, approximate size of the project, URL, anticipated life of the project, how it is going to be distributed);
- any future use you envision (e.g. if you are preparing a CD-ROM for educational use, you might want to consider web delivery too);
- the specific rights you seek (e.g., for use in all classes you teach, or your department teaches; for all editions and

future revisions of the published CD-ROM; for USA, North American, or world rights; for English-language specific languages, or all languages; for print, performance, display, or electronic rights).

You should ask for instruction on the wording of the credit line, the copyright notice related to their material, any other conditions they might have, and the fees that might apply. You should also ask for confirmation that they have authority to grant permission and if not, you should request that they direct you to the appropriate rights-holders. You may request them to warrant that your use will not, to their knowledge, infringe the rights of any other third party.

Make sure to document all your efforts to trace the rights-holders, since if they prove to be untraceable or unresponsive and you decide to go ahead with the project, such documentation could help to prove "good faith best efforts," or "due diligence" if original rights-holders were to initiate legal proceedings at a later date. Such evidence might enable you to argue that the economic impact was limited as the owner(s) made it difficult to license the information.

Licensing schemes

Given the considerable effort and resources that cultural organizations expend in creating digital assets, issues of information control, licensing structures, and royalties deserve careful attention. Strict security measures can constrain and frustrate legitimate users, while rights management tools currently require a significant investment. Additionally, even the best technology-enabled protection schemes can be defeated by persistent hackers. And, while such hacking is illegal under the DMCA, pursuing infringers in court is costly and time consuming.

After considering these issues, organizations normally conclude that it is more sensible to develop an effective business model for the use of digital collections. Many of these business models depend upon licensing, which can be helpful either in obtaining permission to digitize material or in distributing it to others. Site licensing is one model for managing rights in and access to digital resources. Organizations have been experimenting with different licensing schemes for several decades. These schemes usually take into account the number of users and the types of institutions, for example by offering special educational licenses to universities. Licensing provides one fairly secure way of protecting rights and ensures that the holders (or their distributor) can track the users. Another effective mechanism may be to join rights-holders' consortia, as was mentioned in the IPR Management section of Section III: Selecting Materials, which also discusses briefly the issues that consortia raise.

Example Box:

Examples of Licensing Schemes

SCRAN: an example of a licensing scheme

SCRAN, the Scottish Cultural Resources Network (http://www.scran.ac.uk), which has supported the digitization of cultural assets by museums, libraries, and archives, has created a large multimedia resource about Scottish culture and history. SCRAN has developed a Contributor License in which the contributor (e.g. a museum, library, or archive) retains all commercialization rights in the new digitized object, but grants to SCRAN a perpetual, non-exclusive right to use the digital object for any non-profit educational purpose, worldwide. SCRAN passes back to the contributor an agreed payment for any commercial use of the digital object. If SCRAN were to be taken over by any commercial third party, all the educational rights would return to the original contributor. With the User License, SCRAN grants a non-transferable, non-exclusive right to use of the digital objects by the employees and students of the user institution and any member of the public visiting the institution for educational purposes. In addition, the project has devised an extensive range of usage fees, which covers individuals, schools, higher education institutions, libraries, museums, and commercial users and takes into account their size and usage. All resources are digitized at a very high resolution. 'From this archival resource, a network surrogate is created at a lower resolution, to minimize download time. This networked resource may only be downloaded by members of an educational institution licensed by SCRAN. It is protected by an invisible "watermark" (to confirm the resource's copyright status) and "fingerprint" (to identify who downloaded it and when). To avoid any accusations of "entrapment", this information is also clearly shown in banners at the top and bottom of the downloaded image. But any member of the public, anywhere in the world, has access to a thumbnail image of the asset, plus full textual documentation' (Royan 2000).

AMICO: collective licensing

The Art Museum Image Consortium (http://www.amico.org) is an independent nonprofit consortium of organizations with collections of works of art that seek to deploy the digital documentation of those works for educational use. Members are collectively building the AMICO Library as a centralized digital educational resource that is licensed, under subscription, to universities and colleges, public libraries, elementary and secondary schools, museums and galleries.

When subscribing to the AMICO Library, educational institutions select a distributor, such as the Research Libraries Group, SCRAN or H.W. Wilson. Authorized users from AMICO subscribers and members then access the Library through a secure electronic distribution system. The public web site does not allow access to the full Library, but includes thumbnail images

and brief text descriptions of all works. Educational institutions are charged an annual fee to subscribe to the Library, which provides unlimited access to the institution's authorized users for a year.

International perspective

Although cultural organizations operate in an age of instant, global communications, copyright law is still largely earth-bound, the creature of national legislation. In general, countries establish their own copyright laws and enforce those rules within their territories. For example, Country A might provide for a copyright term of life of the author plus 70 years, while Country B may set the term at life of the author plus 50 years. Within this framework of national laws, a number of international treaties (such as the Berne Convention and the Universal Copyright Convention) establish certain universal "minimum standards," harmonize many of the disparate aspects of national copyright laws, and ensure that protection is available across national borders on a non-discriminatory basis.

Nonetheless, cultural organizations may encounter complex legal issues when infringement occurs across national borders. Suppose that a cultural organization located in Country B digitizes (without permission of the author) a copyrighted book written by a citizen of Country A. Since both Country A and Country B are members of the Berne Union, the author from Country A will be entitled to the same rights (and subject to the same limitations) as Country B extends to its own authors.

In general, the law of the country where the infringement took place (and where relief is sought) will govern. Thus, to continue the example, if the law of Country B excuses the infringement under a broad fair use provision, the author may be out of luck even though the law of Country A contains no similar defense. Moreover, in an electronic environment, where material is transmitted from one country to another with the press of a button, determining the place of infringement (and applicable law) itself may be difficult. In analyzing situations involving possible transnational infringements, cultural organizations should begin by posing the following questions:

- in what countries does the institution seek protection?
- what protection is available under that country's national intellectual property laws?
- what treaty provisions may provide protection or facilitate obtaining intellectual property rights in the absence of or as a supplement to local laws? (Shapiro & Miller 1999)

Link Box

Developments to Watch:

- the 2001 'European Directive on Harmonization of Copyright' which tries to harmonize European law: http://europa.eu.int/eur-lex/en/dat/2001/l_167/l_16720010622en00100019.pdf. For a commentary on some of its key elements and how it might affect cultural institutions see the NINCH website: http://www.cni.org/Hforums/ninch-announce/2001/0075.html.
- the Hague Conference on Private International Law that aims to address jurisdiction issues: http://www.cptech.org/ecom/jurisdiction/hague.html.
- the 1994 'Uruguay Round Agreements Act', which, among others, restores copyright in foreign works that have entered the public domain in the U.S: http://www.loc.gov/copyright/gatt.html. Diane Zorich (2000) discusses some of the implications of the Act for the cultural sector in an article that can be found at the NINCH website: http://www.ninch.org/copyright/2000/chicagozorich.html.

Managing IPR risk

Protecting and managing intellectual property rights, and avoiding their infringement in the cultural sector, all involve risk management. When cultural institutions design digital collections, they may well select material in which they do not hold copyright. It is easy to imagine a situation where material an institution would like to use has not entered the public domain, the kind of use it would like to make of the material is not covered by 'fair use', and it has not been able to obtain permission to use the material. For example, a local history museum is creating a new CD-ROM for sale and wants to include digital images of a series of landscapes which were discovered in the attic of the local arts college. These were created after 1980, judging by the buildings that some of them portray, by an artist whose identity is not known to the museum. Despite showing the material to college staff and searching relevant records and making best endeavors (all diligently documented) to contact students who attended the college at that time, the museum does not succeed in identifying the artist and tracing the copyright holder(s). In this case, the staff would have to assess the risk of proceeding and balance the benefits to be derived from using these particular materials for the project, against risks such as negative publicity, costs of litigation, financial penalties that might be awarded, costs of lost human resources (e.g., administrative time), or financial loss to be incurred by having to withdraw the digital resource. In some cases, the staff might conclude that there are important considerations

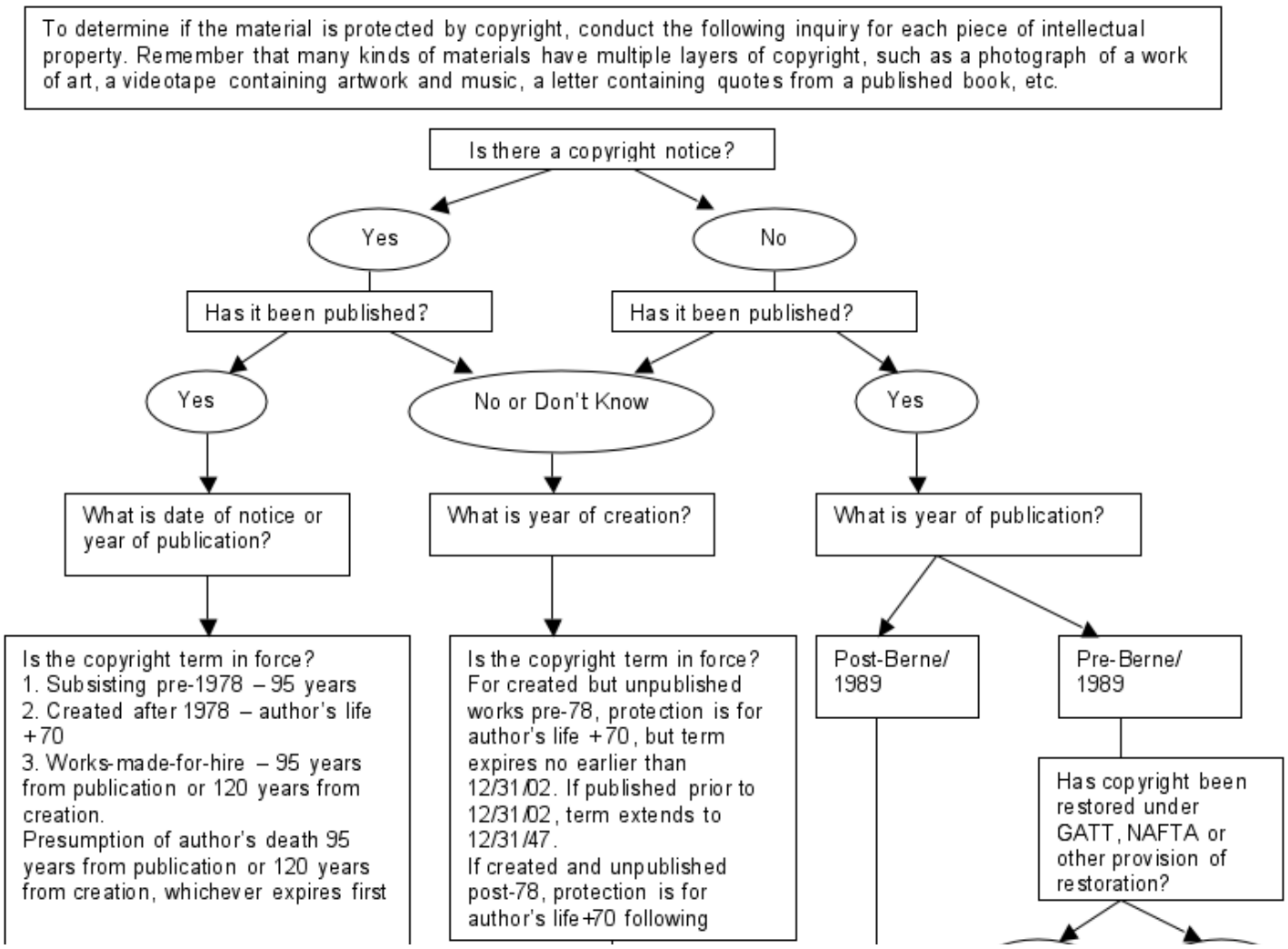
favoring limited use of the material that would counterbalance the risk of infringing the legal rights of the unidentified copyright owner.

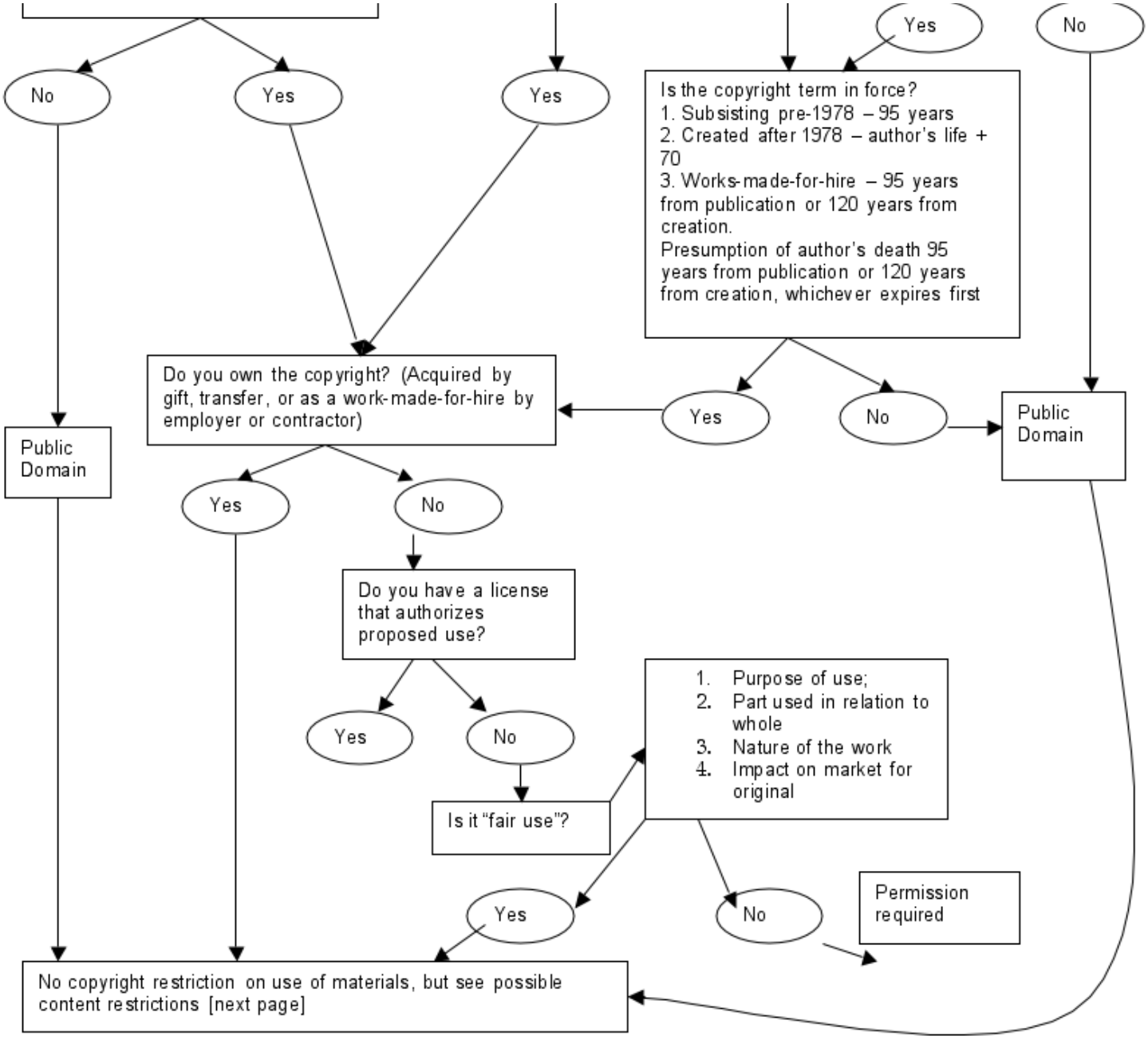
The discussion of Project Management in the Resources section encourages institutions to produce a risk table as part of good project planning. One large category of risk that would be covered in that table relates to rights issues. Here we repeat the advice in that Section, but focus it more on evaluating issues associated with rights.

- What can go wrong? (e.g., What could be the consequences of using the material without specific permission, pursuing a fair use or mitigation of damages argument?)
- For each risk, what is the likelihood of actual risk for the project? (e.g., Has a thorough search to trace the rights holder been made without any results? How can you demonstrate that this search was conducted in a comprehensive manner? Have these efforts been documented?)
- What can be done to avoid the risk or to minimize its impact? (e.g. Have these efforts been documented?)
- If copyright infringement does occur, what would be the impact on the project/program? (e.g., What would be the impact in terms of scale in time, finance or reputation?)
- How will the institution address the risk if it does occur? (e.g. Will it withdraw the material or negotiate with the right holder? Will the institution be able to afford legal expenses?) [3]

As one expert suggests “the test [we] ought probably to apply, is not ‘is this legal?’, but ‘does this harm any one?’. If an action does not harm the financial interests of another institution or individual, it is probably safe to take it” (Royan 1998). While there may be merit in Royan's argument, especially where your institution can demonstrate that it has acted with due diligence to identify the holder of the rights without success (see above), cultural institutions must be aware of copyright regulations and act prudently to avoid explicitly infringing them. So before investing considerable effort in digitizing your collections, you should address the questions outlined in the following decision tree diagram, where we summarize the questions raised throughout this section:

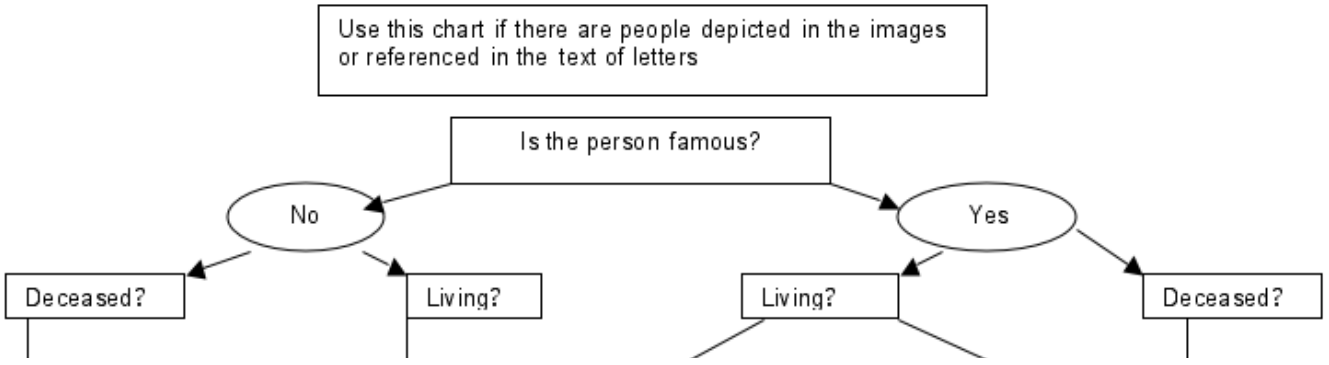
1. REPRODUCTION OF MATERIALS ON WEB SITE

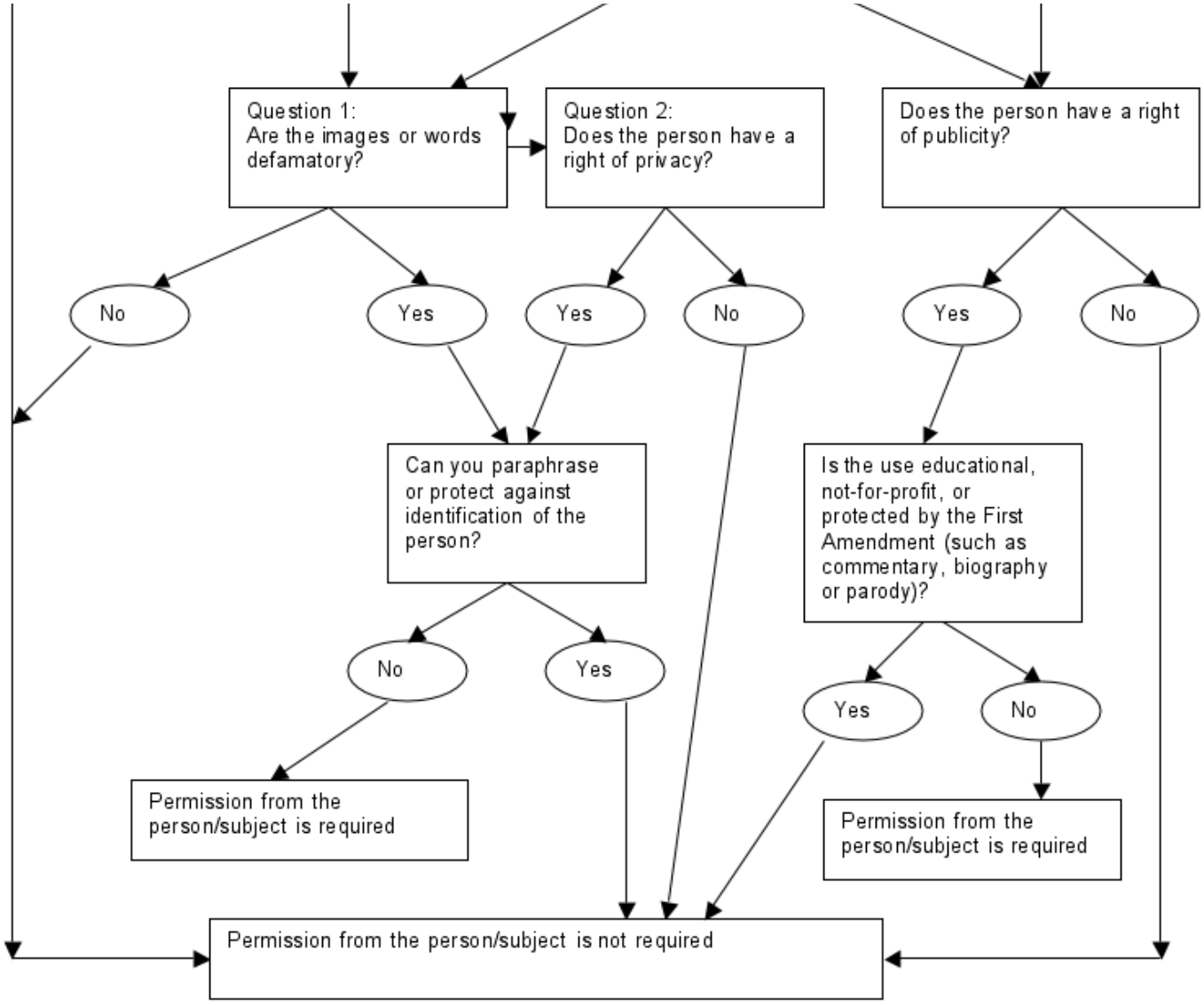




Copyright, Lauryn G. Grant; reproduced with permission

2. CONTENT RESTRICTIONS FOR REPRODUCTION OF MATERIALS ON MUSEUM WEB SITE





Right of Privacy - the right to be left alone, not portrayed in a false light, not have private facts publicly divulged, not have life intruded upon, and not have name or image misappropriated. The right applies only to living persons.

Defamation - the right to prevent false and malicious comments about you from being published.

Right of Publicity - the right of celebrities to control the use of their name, image, likeness, signature and voice. The scope of the right and how long it lasts are subject to state law.

Trademark - cannot use trademarks or logos for a commercial purpose or if the use creates confusion as to the source of the product or service.

Copyright, Lauryn G. Grant; reproduced with permission

Link Box:

Rights Management Resources

There is a wealth of available information on rights management issues on the Web. Good starting points include:

- ***U.S. Copyright Office***, Library of Congress: http://Lcweb.loc.gov/copyright
- ***American Library Association*** Washington Office: Copyright & Database protection page: http://www.ala.org/washoff/ip.html
- Shapiro, Michael & Miller, Brett 1999. *A Museum Guide to Copyright and Trademark.* American Association of Museums. Highlights of the book are available on the AAM site: http://www.aam-us.org/resources/reference_library/mus_guide_copyright.cfm.
- McCord Hoffman, Gretchen. 2001. *Copyright in Cyberspace: Questions & Answers for Librarians*. Neal-Schuman Publishers.
- The ***Digital Millennium Copyright Act*** of 1998 (http://www.loc.gov/copyright/legislation/dmca.pdf) implemented the treaties signed in December 1996 at the World Intellectual Property Organization (WIPO) Geneva conference, but also contains additional provisions addressing related matters (some of which were opposed by many scientists, librarians, and academics). For a discussion of the key points, see the site of the UCLA Online Institute for Cyberspace Law and Policy: http://www.gseis.ucla.edu/iclp/dmca1.htm.
- The ***Digital Future Coalition*** (http://www.dfc.org) is a group 'committed to striking an appropriate balance in law and public policy between protecting intellectual property and affording public access to it'. The Coalition brings together non-profit educational and cultural organizations together major commercial associations from the computer, telecommunications, and network access industries.
- Sinacore-Guinn, David, 1993. Collective Administration of Copyrights and Neighboring Rights: International Practices, Procedures and Organizations (Little Brown and Company). An excellent compilation on how rights management does and should work.

Rights of privacy

The massive availability of digital resources has led to the exposure of much previously unpublished material, including personal, commercial or sensitive information. In the selection stage, but also when considering public access to digitized material, the project staff should ensure that the material is handled with responsibility, sensitivity, and care.

Projects that involve the digitization of personal data should be aware of the data protection legislation of the country where the project is based. The different notions of privacy in different nations' laws will continue to have an uncertain impact in the cultural heritage field. Some general key guidelines are:

- Personal data should be obtained lawfully;
- Personal data should be collected, kept, and used for specified purposes;
- Rights of access and of processing personal data should be clearly specified;
- Where users are provided access to personal data, your institutions should maintain an audit trail to track who has accessed the data and how and when they used it.

In the U.S., privacy on the Internet is largely self-regulated. Increasingly, however, companies are shunning sites and associations that do not follow basic privacy safeguards. The Direct Marketing Association has gone so far as to expel members who do not follow basic privacy safeguards, for instance those who do not post privacy statements. The industry is attempting to avoid federal legislation in this area through self-regulation. Although the Federal Trade Commission (FTC) has so far not recommended any Internet privacy legislation except with respect to children, a provider of information — who collects personal information should disclose the following information: why it is being collected; how it will be used; what mechanisms are in place to protect its confidentiality, ensure its quality, and guarantee its integrity; what impact users' decisions to provide or withhold information will have on the services they can access; and, what redress individuals have if inaccurate information is disclosed. If you are a provider of information and collect any information about your users you should have a privacy policy on your website and it should be clearly labeled and easily accessible to users.

Institutional policies can be very useful for navigating this complex area of information privacy. For example, the information privacy policy of the Science Museum of Minnesota (http://www.sci.mus.mn.us/museum/identity/Privacy.html), covers, among other things, information collected by the Museum and submitted by third parties, personal and other information collected on its website, the use and sharing of information by the Museum (including use of postal addresses, telephone numbers, and emails and how to opt-out), the use of cookies, and children's privacy on the web (with reference to the Children's Online Privacy Protection Act).

Example Box:

In the case of the Survivors of the SHOAH Visual History Foundation (VHF), the copyright in the interviews with the Holocaust survivors has been assigned to the VHF. Each interviewee signed a release that granted the Foundation the rights to distribute the testimonies for educational purposes. VHF recognizes its duty to protect the sensitive material contained in the testimonies from misuse and misrepresentation. This duty is reflected in the approach that the VHF takes to the distribution of the testimonies. The VHF has identified four categories of use:

- ***Foundation-authenticated use***, by users who are using the archive within the Foundation itself, who are known to the VHF and with whom they have a direct contractual relationship. These users apply to use the archive and sign a non-disclosure document first to conduct research. If they seek to publish, they return to the Foundation for a license for permission to publish and/or to obtain broadcast quality video for use in an exhibit or documentary.
- ***Location-authenticated use***, where the VHF knows who and where the users are. This use will be through the interfaces at identified museums, such as the Holocaust Memorial Museum in Washington, D.C. or the Museum of Tolerance Simon Wiesenthal Center in Los Angeles.
- ***Location-unauthenticated use***, where the VHF knows where but not who the user is. This may be an exhibit in a museum or restricted use of the catalog and finding aids.
- ***Globally-unauthenticated use***, where the VHF knows neither who nor where the user is; potentially the Internet at large. This level of use is not being permitted by the VHF yet, as the staff believe that much work will have to be done on the authenticated users first in order to understand how the materials might be used and how they should be accessed.

Other intellectual property concerns

The bulk of this section has dealt with copyright, although as we have just seen privacy concerns are important as well. Even when a work you plan to use is not protected by copyright law or does not have privacy issues associated with it, other intellectual property rights may come into play. The material may be protected by other laws, such as publicity, patent, trademark, and trade secret laws.

Databases

The availability of information is of paramount importance to maximizing the potential of the Internet. Legal protection of databases is therefore an important and timely issue. In some countries, databases are protected by copyright as original compilations of information, where their creation has involved effort. Until the ruling in the U.S. case of FEIST PUBLICATIONS, INC. v. RURAL TELEPHONE SERVICE CO., (499 U.S. 340)[4] in 1991 that only original creations were afforded copyright protection, some had assumed that creations which depended solely upon ‘the sweat of the brow’ were afforded protection. In this instance the Court held that as a telephone directory was simply an alphabetic listing it did not contain sufficient creative expression to be eligible for copyright protection. This judgment has been widely interpreted to imply that in the U.S. there is little copyright protection of databases of facts which exhibit no creative originality. This is in line with most of continental Europe. In order for a database to warrant copyright protection, its author must have made some original effort in the collection, selection, and arrangement of material (e.g. by indexing terms or adding keywords). This protection is irrespective of whether the individual contents of the database are original and therefore copyrightable in themselves or include factual data (where the database protection will not prevent an individual from extracting them, short of copying the selection and arrangement of the database as a whole). Currently in the U.S., if the database is unprotected by copyright law, the contents of the database may be copied unless such acts are prohibited by contract or license. In Europe, however, the Database Directive adopted by the European Parliament in 1996 includes a right that prohibits (for fifteen years from the date of the database’s creation) the extraction or reutilization of any database in which there has been a substantial investment in either obtaining, verifying, or presenting the data contents, without any requirement for creativity or originality. Databases are thereby afforded some element of sui generis legal protection. (See the Copyright office’s 1997 Report on Legal Protection of Databases at http://www.copyright.gov/reports/ for a good discussion of contractual practices for accessing data.)

Watch this space: This is a hotly debated issue at the moment in the United States.

Registration

Registration provides a mechanism to strengthen the protection that copyright provides. Even though it is no longer necessary to use the copyright symbol © or print a copyright notice in order for your materials to be afforded copyright protection, doing so will ensure that the court will not accept a defendant's 'innocent infringement defense'. Similarly, while registering your materials with the Copyright Office is no longer a requirement, registration offers several benefits which are otherwise lost.

These include the ability to bring an infringement case to court, to prove certain facts in a lawsuit, and to claim statutory damages and attorney's fees. Registration material is available in the U.S. at the Copyright Office, Library of Congress (http://www.copyright.gov/register/index.html).

The preparation of effective copyright notices depends upon knowledge of legal frameworks not only in the U.S., but in other countries. For example, according to the Buenos Aires Convention, in order for your material to gain copyright protection in Bolivia and Honduras, you must include the phrase “all rights reserved”. While preparation of copyright notices is best done with appropriate legal advice, Benedict O'Mahoney's copyright website includes valuable guidance on how to draft such notices (http://www.copyrightwebsite.com/info/notice/notice.asp).

Technologies for copyright management and protection

Throughout this section we have discussed the need to strike a balance between protecting intellectual property and providing public access to it. In an effort to meet both these objectives there has been an increase in the design of electronic mechanisms which can mark or tag digital cultural assets and trace their usage. In addition to these technical mechanisms, the use of appropriate metadata can play an important role in making users aware of who owns the rights in material, the kinds of uses to which the users are allowed to put the materials, and the conditions that govern different uses (e.g. free to look at on your screen, but you must pay a fee if you wish to forward it to someone else). The encoded metadata might include a description of the object, details of the rights owner, as well as a unique registration number.

Digital Object Identifier (DOI)

The ***Digital Object Identifier*** (DOI), a system which has been proposed by the International DOI Foundation (http://www.doi.org) in conjunction with the Association of American Publishers and the Corporation for National Research Initiatives was launched in 1997 and is used by several publishers. It provides a unique identification number to digital objects (as one metadata element), allows rights holders to link to the users of the material, and enables automated copyright management for all types of media. By clicking on a DOI, the user is connected to a central directory where the Web address used by the publisher is located. This address is maintained by the publisher and may contain the content itself or further information on how to obtain the content and means of linking users to rights holders. As this might also contain a description of the object, it is important that DOI developments take into account the Dublin Core scheme and that the two communities work together towards harmonization of the two models (Bearman 1999).

Despite active research, currently available technologies for protecting assets have limited effectiveness and little general acceptance. They all aim to provide a mechanism recording and linking rights information to the digital object that is indelible and unmodifiable but perhaps also updateable and able to 'follow or accompany' the digital object when copied. Research for this Guide showed that most projects have chosen to not provide high quality archival or even medium quality resources online (e.g. Images of England, http://www.imagesofengland.org.uk), while they wait for digital rights protection technologies to become more effective, standardized, widely used, and affordable.

Link Box:

Some examples of advanced technologies for management and protection of digital assets include:

- ***steganography***, which hides the information under a stegoimage, a cover image where the original is embedded (Johnson & Jajodia 1998, http://www.jjtc.com/pub/r2026a.htm; see also the infosyssec site on Cryptography, Encryption and Steganography, http://www.infosyssec.org/infosyssec/cry2.htm)
- ***cryptography*** which encrypts or scrambles the information so that it cannot be understood without a decoder, a pre-defined decryption key (for an introduction to cryptography, see the page by SSH Communications Security, http://www.ssh.fi/tech/crypto/intro.cfm)
- ***digital wrappers***, protection systems which place digital content inside secure 'containers' which require a helper application and a key code to be opened. For an example see InterTrust's DigiBoxes, http://www.intertrust.com.
- ***digital watermarking***, which is a digital signal or pattern inserted into a digital image and can be either visible or invisible. An early example includes the visible watermarks designed by IBM which were inserted in the images of the Vatican Digital Library Project (Mintzer et. al. 1996, http://www.research.ibm.com/journal/rd/mintz/mintzer.html). Although visible watermarks discourage theft and misuse, they can interfere with studying the original and decrease

the overall quality of the digital object, so there seems to be a preference for invisible ones in the cultural and education sector. Even the most sophisticated digital watermarks however, can be overcome through overmarking, adding noise, removing the watermarking, and counterfeiting. For example, the Secure Digital Music Initiative (SDMI, http://www.sdmi.org) digital watermarks were defeated within three weeks. For more information on digital watermarking and steganography, see Fabien Petitcolas' page at the University of Cambridge Computer Laboratory, http://www.cl.cam.ac.uk/~fapp2/steganography/.

---

[1] Section 107 of U.S. Copyright law: http://www.loc.gov/copyright/title17/92chap1.html#107

[2] Section 106a of U.S. Copyright law: http://www.loc.gov/copyright/title17/92chap1.html#106a

[3] For further information on risk management see: http://www.utsystem.edu/ogc/intellectualproperty/riskmgt.htm

[4] See http://www.law.cornell.edu/copyright/cases/499_US_340.htm for details of this case.

table of contents previous chapter next chapter

valid xhtml 1.1
abp~03/03

V. Digitization and Encoding of Text

Introduction

Digitized text is an important component of many cultural heritage projects, but the question of how to digitize it—what format, how much detail, what user activities to support—is a complex one. As with the digitization of other kinds of materials, you must consider a number of factors including the nature of the original material, the purpose of the digitized version, and the availability of relevant expertise, technical support, and funding.

It is also important to be aware of the purposes and limitations of the various digital formats available, to be sure not only that they suit your current goals and requirements, but also that they do not restrict your options in the future. Digital text formats vary considerably in the ease with which they can be converted to other formats, and in the variety of output methods they support. Proprietary systems such as word processing or page description formats (e.g. Microsoft Word, PDF) may be powerful and convenient tools for creating printed output, and may also allow for web publication of the results (via a "save as HTML" function), but if you need to move your data to another software platform you risk losing formatting and other information. Because such systems depend on the existence of proprietary software—whose licensing terms and very existence cannot be counted on over the long term—they are unsuited for archival purposes or for the creation of durable cultural resources.

In the past ten years, there has been a rapid growth in standards-based methods of text digitization using Standard Generalized Markup Language (SGML) and more recently its derivative, Extensible Markup Language (XML). These approaches avoid the problems of proprietary software, offering data longevity and the flexibility to move from platform to platform freely. There are now increasing numbers of tools for creating, editing, publishing, and manipulating textual materials encoded in this way, and this trend is likely to continue. Our treatment of digital text will therefore focus on standards-based methods of text digitization, which offers the best long-term solution to the needs of projects creating digital cultural heritage collections.

It may be useful to establish some basic terminology at the start. By "digitization" we mean any process by which information is captured in digital form, whether as an image, as textual data, as a sound file, or any other format. When speaking of the digitization of documents, the term may refer either to the capture of page images—merely a picture of the document—or to the capture of a full-text version, in which the document is stored as textual characters. In its most minimal "plain-text" form, a full-text version of a document may be simply that: the text of the document expressed as ASCII or Unicode characters and nothing more. Unlike a page image, such a document can be searched for particular words or phrases, but does not convey any information about the original appearance or structure of the document. An "encoded" version of the same document will include additional information or "markup" of various kinds, expressing the document's structure, its formatting, or other information its creators wish to capture. Although strictly speaking, we can use the terms "markup" and "encoding" to refer to a wide range of added information—including word processor formatting codes or encryption—these words are now most frequently used to refer to SGML or XML markup. And although this kind of markup is usually applied to full-text documents, it is also possible to embed page images in an SGML or XML-encoded document structure, or to pair images with encoded information such as subject keywording, publication data, or administrative metadata.

Definition Box:

Definitions

Page Image: A digital image of a page of text, captured by a scanner or digital camera, and expressed as a set of pixels in a format such as JPEG or TIFF.

Encoding, markup: In this context, the process of adding information to a digital text by including markup

(usually SGML or XML) which explicitly identifies structural and other features of the text. The term “markup” refers to the added information. In a broader sense, encoding may refer to any kind of added information or algorithmic transformation which, when applied to a data file, enables it to perform some special function.

Keying: A process by which a person manually types, or ‘keys,’ text from source page images, original printed materials, photocopies, or microforms.

OCR: Optical Character Recognition, a process by which software reads a page image and translates it into a text file by recognizing the shapes of the letters with various levels of accuracy that are difficult to predict. OCR-generated text tends to be described as either “uncorrected (or raw)” or as “corrected.”

Metadata: Strictly speaking, any data that is about other data; in this context, more specifically, the term usually refers to information describing a data file or document (for instance, publication information, revision history, data format, rights status).

SGML: Standard Generalized Markup Language, an international standard (ISO 8879) since 1986, is a metalanguage which can be used to define markup languages.

XML: Extensible Markup Language, a subset of SGML which was published as a W3C recommendation in 1998.

TEI: The Text Encoding Initiative, an international consortium that publishes the TEI Guidelines for Electronic Text Encoding and Interchange, an SGML- and XML- compliant encoding language for the capture of literary and linguistic texts, widely used in the scholarly and cultural heritage communities.

EAD: Encoded Archival Description, an SGML- and XML-compliant encoding language used for the capture of archival finding aids, widely used in the library and archival communities.

METS: Metadata Encoding and Transmission Standard, an XML-compliant standard for encoding a variety of metadata about digital library objects.

Having selected the materials you wish to digitize, (and having decided whether your text digitization strategy will include producing page images) there are several methods to choose from.

Page images are produced by scanning the text. Optical character recognition software reads such an image and creates a full-text version of the document by identifying individual character shapes and translating them into actual letters. Note that there are many factors affecting the accuracy of the OCR process, including the contrast of the original document, the fonts used, etc. Alternatively, if you have decided that page images are not needed, you can key text directly from a variety of sources. The formats available, and the handling policies pertaining to them, will dictate whether keying will be done from original materials, photocopies, or microforms. You can have the document keyed by project personnel or by a data capture service. Markup can be added during the keying process or subsequently. Increasingly, data capture services are willing to cater for specialized encoding needs, even using complex markup languages like those in the guidelines of the Text Encoding Initiative (TEI) and Encoded Archival Description (EAD); see the more detailed descriptions below and the URLs at the end of the chapter. It may be also possible to automate the encoding to some extent; how much will depend on several factors, including the consistency and predictability of your data, the level of encoding you require, and the amount of time and ingenuity you are willing to spend developing automated encoding tools. Some tools for this purpose exist already, and it is worth checking with existing encoding projects to see whether they may have something which will accomplish what you need. Finally, you may be able to obtain a digitized version of the materials you need from another project, such as the Oxford Text Archive or the University of Virginia Etext Center, either in plain text or with basic text encoding and metadata.

A few concrete scenarios may help to illustrate the differences between these various approaches, and the kinds of project goals each one is best suited to support. First, consider a fairly straightforward example: a project working with a collection of Victorian novels and poetry for which it owns copies of the materials to be digitized (the Victorian Women Writers Project, http://www.indiana.edu/~letrs/vwwp/, is an example of such a project). Since the materials are fragile and of historical interest, capturing page images makes sense both as a form of digital preservation, and in order to give readers access to an image of the original materials for research purposes. (However, given the uniformity of these texts and the general unremarkability of the pages, one can also imagine deciding against this step, particularly if the project did not own copies of the materials.) In addition,

since these are literary documents, researchers will want to work closely with the text, so capturing a full-text version also makes sense. Adding TEI markup will enable readers to perform complex searches which take into account the genres and local textual structures which may be of interest. Since the page image has already been digitized, and since the materials in question are typographically regular, OCR is a good option for capturing the text, but since the project's audience is a scholarly one, careful proofreading will also be necessary to ensure that requirements for accuracy have been met. The question of how deeply to mark up the documents requires careful balancing of costs and desired function. Basic TEI markup can be applied nearly automatically using scripts, but more detailed encoding requires additional staff, training time, and additional review for encoding consistency. The additional benefit to researchers would have to be substantial to justify this cost.

Consider another project in which a large quantity of heterogeneous materials must be digitized quickly, and made accessible for political and historical research (for instance, the Forced Migration Online collection at the Refugee Studies Project Centre at Oxford, http://www.rsc.ox.ac.uk/TextWeb/fmo.html). Retrieval is acknowledged as an important goal, but at the same time the nature of the individual documents does not seem to warrant complex markup; most users want to find documents relevant to their research, but are not interested in the texts' internal structure. Furthermore, given the volume of documents and the urgency of the project, costs and speed are important factors. The project's designers therefore decide to digitize all the documents as page images; since the documents vary so much and are likely to be unfamiliar, it is important for readers to see an image of the original. In addition, to support detailed retrieval at the topic level, they decide to OCR the text, but because of the volume and cost issues, they choose not to proofread each document. Since users are searching through such large volumes of material, the error rate of uncorrected OCR still provides acceptable accuracy when used only for indexing and query purposes. Users will see only the page images, not the full-text version, so the errors will not be visible. In addition, the project creates basic metadata for each document, to allow for accurate tracking and high-level retrieval.

These two kinds of projects have arrived at what look like fairly similar approaches (page images, text captured with OCR, and metadata), though for very different reasons. Consider now a third project, whose materials are medieval manuscripts to be digitized for literary and linguistic study. (See, for example, The Electronic Beowulf project, http://www.uky.edu/~kiernan/eBeowulf/guide.htm.) As before, there are likely to be important reasons to include images of manuscripts, but their uniqueness and fragility magnify the legal and logistical challenges. Because of the nature of the research to be supported, a full-text version is essential, but OCR is probably out of the question. Instead, the project needs to choose an appropriate method for keying in the text. For medieval manuscripts, this may require a scholar familiar with the manuscript to perform or oversee the transcription. Some materials may be readable enough by a non-expert to allow for keyboarding by a data capture service, or by locally trained encoders who are not subject specialists. Finally, to support literary and linguistic analysis, the project needs to use a detailed markup scheme such as the TEI, which will allow for the capture of features such as regularized or lemmatized readings of words, morphological and syntactic information, prosodic structures, textual variants, ambiguous readings, and so forth. With page images linked to this level of encoded full text, scholars can compare versions of the manuscript, check difficult readings against the image, search for particular linguistic features and compare their occurrence from poem to poem or from manuscript to manuscript. The collection is a full-fledged scholarly tool of extraordinary power, but it also requires considerable resources and expertise to create.

Projects planning to capture a full-text version of the document should also refer to the sections on OCR (Optical Character Recognition) and Keyboarding in Section VI, Images, in addition to the sections below.

Character encoding

Even a plain-text document without markup contains some very basic encoding of the character information in which the document is expressed. There are two principal character encoding systems which deserve brief discussion here: ASCII and Unicode.

ASCII—the American Standard Code for Information Interchange—was proposed by ANSI (the American National Standards Institute) in 1963, and finalized in 1968. It was designed to facilitate compatibility between various types of data processing equipment. ASCII assigns the 128 decimal numbers from 0 to 127 to letters, numbers, punctuation marks and common special characters. This is commonly referred to as the ‘low ASCII character set’. Various extensions to the ASCII character set have been created over the years to assign the decimal numbers between 128 and 255 to special, mathematical, graphical and non-Roman characters. If you wish to provide texts in a ‘plain text format’, i.e. with the file extension “txt”, then you must use the low ASCII character set. The extended ASCII character set has limitations that do not apply to the lower set and must be

used with more caution. There is also more than one extended character set (IBM and Microsoft each have their own) and this diminishes interoperability.

There is some argument to be made for always providing a plain text format without markup, whatever other encoding scheme is used, especially in the creation of electronic corpora. There is not much extra work involved in making the texts available in this simplified form, especially as plain ASCII texts are often the most common starting point for creating other encoded text. Certainly encoded texts (for instance, SGML/XML, COCOA, or other formats) should be created and stored in plain ASCII rather than in a proprietary system such as a word-processing format, to guarantee their longevity and platform-independence. Projects such as the Thesaurus Musicarum Latinarum (TML), where Latin musical notation has been encoded using the low ASCII character set, demonstrate that encoding complex features, while maintaining a high degree of interoperability and longevity, can be achieved with ASCII characters. This high level of interoperability is particularly important for the TML because its small but highly dispersed user base uses a wide variety of hardware and software. Furthermore, the small file size of ASCII text files means the output of the project can be distributed quickly and easily.

Unicode is an international standard (implementing ISO 10646) which was developed to overcome problems inherent in earlier character encoding schemes, where the same number can be used for two different characters or different numbers used for the same character. Unicode assigns a unique number to each character; in addition, it is multi-platform, multi-language and can encode mathematical and technical symbols. It is especially useful for multilingual texts and scripts which read right to left. Originally there were 65,000 characters available; now there are three encoding forms which can be used to represent over 1,000,000 characters.

Link Box for character encoding:

Unicode: http://www.unicode.org/

ASCII: http://www.asciitable.com/

Other formats which were used in the past are worth remembering:

- EBCDIC = Extended Binary Coded Decimal Information Code, the 8-bit character coding scheme used by IBM mainframes and minicomputers.
- Multicode: (see Mudawwar 1997)

Text markup

Text markup is a fundamentally interpretive activity, in ways that are both powerful and challenging. With a well-designed encoding language, you can express the facts and insights which are most useful for your purposes in working with the text. However, the interpretive quality of text encoding should not be taken to mean that encoded texts are purely, or merely, subjective creations. The most widely-used humanities text encoding languages such as TEI or EAD have been developed by particular communities whose needs and methodological assumptions they reflect.[1] These languages are designed to allow the representation of the significant information these communities want to capture. This information includes not only basic facts—or assumptions so deeply shared that there is no disagreement about them—but also significant interpretive statements, such as those of critical editing or archival description. One important role of the encoding language is to enable such statements to be made in a rigorous and consistent way, according to the practices of the community within which they are to be used. Text encoding thus makes it possible to bridge the gap between local research and insight and the discourse of the larger community, and to articulate interpretative statements in a way that is broadly intelligible.

The example given here shows a historical letter encoded using the guidelines of the Model Editions Partnership (MEP, http://mep.cla.sc.edu). As with all SGML-encoded documents, each textual element is enclosed within tags (set off by angle brackets) which mark its beginning and ending. The encoded text begins with a header (here abbreviated for simplicity) which contains the document metadata. Following this, the body of the document (encoded as <docBody>) contains first a heading, dateline, and salutation, and then the main text of the letter encoded as a series of paragraphs. At the end is a postscript with its own dateline. The elements and their grouping reflects the interests and research needs of documentary historians, for whom the MEP encoding

scheme is designed.

Example Box:

A Sample Fragment of an SGML-Encoded Text (some encoding omitted for clarity)

```
<doc>
<mepHeader> ... </mepHeader>
<docBody>
<head>To <addressee>Martha Laurens</addressee></head>
<dateline>
<place>Charles Town</place>, <date>August 17, 1776</date>
</dateline>
<salute>My Dear Daughter</salute>
<p>It is now upwards of twelve Months ...</p>
...
<p>You will take care of my Polly too ...</p>
<signed>your affectionate Father</signed>
<ps>
<dateline><date>19th</date></dateline>
<p>Casting my Eye over ...</p>
</ps>
</docBody>
</doc>
```

The balance between disciplinary constraint and local expression is managed in SGML and XML through the document type definition (DTD), which is a formal statement of the tags permitted in the encoding system (i.e. what textual structures it is capable of naming) and how they may be nested. Different DTDs handle this balance very differently, depending on their intended purposes. Structurally, DTDs may be quite strict, specifying with great precision the order and nesting of tags, or they may be quite lenient, allowing greater latitude to the encoder. Similarly, in their definition of a tag vocabulary, DTDs may provide for very nuanced distinctions between features, or they may use fewer, more generally defined elements. Finally, DTDs may be constructed in an attempt to anticipate and codify all the possible encoding situations in the domain they cover, or they may instead provide methods for the encoder to handle unforeseen circumstances. All of these approaches have their potential uses, and in choosing or designing an encoding system it is essential to understand the nature of your own material and goals, so that you can choose appropriately. Although a structurally strict encoding system may seem to limit one's options unnecessarily, in fact such a system can be valuable in constraining data entry and in ensuring that identically structured documents are all encoded alike. A more lenient DTD is superficially easier to work with, but through its flexibility it also opens up the likelihood of inconsistency and of time-consuming debate about which option to choose in a given circumstance. A DTD with a rich lexicon of tags may be essential for describing certain kinds of textual features in detail, but is an encumbrance when only a simple encoding is required.

DTDs and the encoding systems they represent can often be adapted by the individual project to suit local

needs, and this is particularly true (and may be particularly useful) in the case of large, multi-genre DTDs like the TEI. However, one important function of the DTD is to allow for comparison and interoperation between collections of encoded data. By changing it, you in effect secede from the encoding community of which that DTD is the expression, and you diminish the possibility of using common tools for analysis, display, and retrieval. Some encoding systems are designed to be extensible; for instance, the TEI provides an explicit mechanism by which individual projects may define TEI-conformant tag sets which are adaptations of the TEI encoding scheme. When done with care—and preferably in concert with other projects with similar needs—such adaptations may improve an existing encoding system while avoiding the disadvantages described above. For projects dealing with highly idiosyncratic data, or projects attempting to capture features for which no encoding system exists, adaptation may be simply unavoidable. In such cases, you should be prepared to think through your changes carefully and document them thoroughly.

SGML and XML

Although there are a number of encoding systems which have been developed for humanities computing use over the past few decades—many of them still in use—Standard Generalized Markup Language (SGML) and its derivative, Extensible Markup Language (XML) deserve particular attention here both because they are so widely used and because they should be. As international standards,[2] they receive a level of attention from software developers and from the standards community which guarantees their comparative longevity, and because they are non-proprietary, they can be used to create archival collections which are free from software or hardware dependencies, and hence less prone to obsolescence.

Strictly speaking, SGML and XML are metalanguages: systems for defining encoding languages. Because they provide a standardized method for specifying things like how a tag is delimited or how its structural definition is written, software written to this standard can be used with documents encoded with any SGML or XML-conformant encoding language, regardless of the particular tag set or the kinds of data they contain. The most significant text encoding systems for cultural resources—TEI, EAD, CIMI, METS, and others—are all written in SGML and XML, as is the ubiquitous HTML. There also exist SGML and XML versions of data standards like MARC.

The advantages of SGML and XML, as suggested above, stem partly from their status as international standards. In addition, because this kind of encoding allows for the complete separation of structure and presentation, SGML/XML-encoded documents can be repurposed or used as a base format from which to derive specific versions for different purposes: word-processing files for printing, HTML for delivery on the web, Braille output for the visually disabled, and so forth. SGML/XML encoding is particularly valuable for the kinds of cultural heritage work covered in this report, because it permits the description of the text's constituent parts in terms which are meaningful for retrieval and intellectual analysis. We might express this whimsically by saying that an encoded document “knows” what its own parts are in the same way that a scholar or reader does: concepts like “heading” and “quotation” and “poem” and “author” are accessible as primary terms of analysis (assuming they are part of the encoding language used). Furthermore, unlike some of the earlier encoding languages that were designed to avoid verbosity at all costs, SGML/XML encoding is actually fairly easy to understand once the eye becomes accustomed to seeing the tags. Encoding languages like TEI and EAD use tag names which are expressive of their function—<note>, <author>, <name>, <list>, <quote>, and the like—and because they represent the ideas people actually have about documents, they quickly become intelligible even to the untrained reader.

The disadvantages of SGML and XML have largely to do with their formalism as data structures. Because of their requirement that all documents be expressed in the form of a hierarchy or tree of nested elements, they are not ideal for representing truly non-hierarchical materials (for instance, sketchbooks). Although XML cannot simultaneously represent multiple hierarchies in the same document (and SGML can do so only with great difficulty), these occur so frequently that systems have been developed to handle most ordinary cases, and in practice this is not usually an obstacle to the use of SGML or XML, only a design consideration.

While SGML in its many applications—HTML, TEI, EAD, and others—is widely used by cultural heritage projects, we are entering a period of transition where XML is becoming more widespread. XML is in effect a streamlined version of SGML, in which some features of SGML which make it unnecessarily complex to implement have been eliminated.[3] Like SGML, XML is a metalanguage which can be used to define specific encoding languages, and all of the encoding languages discussed here are now available in an XML version. However, rather than abandoning SGML in favor of XML, projects seem to be using both. With the growing availability of XML software for publication and browsing of online documents, XML has become a central

component in the delivery of cultural heritage materials, and will only become more so in the future. Several factors are particularly significant in this shift:

- because XML tools are easier to write, there is already significantly more software for processing, publishing, and managing XML data, and these tools are not confined to the business and industrial markets. With the rise of XML, cultural heritage projects will thus find an increase in affordable software that meets their digital publication needs. One extensive though technical resource for information on these kinds of tools is the Cover Pages, maintained by Robin Cover (http://www.oasis-open.org/cover/sgml-xml.html). The TEI also maintains a list of SGML and XML tools at http://www.tei-c.org/Software/.
- publication of XML is already simplified by the availability of tools to transform XML-encoded documents from one encoding scheme to another (e.g. from TEI to HTML), but in addition web browsers are now being developed which can read XML documents and style sheets, and hence can display documents from encoding systems other than HTML.
- XML makes it easier for users to use local style sheets (XSL and CSS) to display information the way they want it. It also makes it possible for document publishers to derive multiple encoded versions of a given document from a single master XML file, using XSLT stylesheets. This feature makes it much easier than before to create and manage versions adapted for different kinds of user needs (for instance, the visually impaired) or different delivery strategies.

Unlike SGML, XML documents do not require a DTD. For publication purposes, the lack of a DTD is not much of a concern as long as the document is well-formed (that is, as long as its elements all nest within one another and as long as they all have start-tags and end-tags). An XML style sheet (XSL) can still format the document and browsers can still display it without checking against a DTD to see that the document is valid. However, for production purposes, working without a DTD is not advisable, since it makes it impossible to check documents for consistency of encoding. An alternative to DTDs is XML Schemas, which are another way to specify and validate the structure of your documents. XML Schemas are currently under review by the W3C, and offer some advantages over DTDs, such as typing of elements, grouping of declarations, and inheritance of properties. For more details, see http://www.w3.org.

Developing a Document Type Definition

One of the first steps for new text encoding projects is to identify the encoding system and specific DTD to be used. As suggested above, this process also involves articulating the project's methodological commitments and audience, as well as its more general area of focus. There are a number of encoding languages that fall within the domain of the cultural heritage community, but their range of usefulness does not overlap by much. The TEI DTD is primarily intended for full-text encoding of humanities and scholarly texts; the EAD DTD addresses the needs of the archival community for encoding finding aids and guides to collections. At this point, projects that deal primarily with artifacts, works of art or audio-visual material are far less well served.

It is probably clear by now that if you can use an existing DTD for your materials, you probably should. However, if your material falls outside the realm of existing encoding systems, you may need to develop one yourself. Projects such as the Oriental Institute of Chicago, whose text material does not fall within the broad western text canon around which TEI and the majority of DTDs have been designed, must either develop their own or await developments from others. While simple DTDs are easy to create, developing an encoding system that will fully represent your materials may require considerable research and time. The complexity of the task increases with the heterogeneity of the collection and the level of detail you wish to represent. For a project with a large number of texts with variable structures and features, it can take many years of development, application and refinement to produce a DTD that meets all of its requirements. This is by no means an impossible task, and important projects like the William Blake Archive (http://www.blakearchive.org/public/about/tech/index.html), The Orlando Project (http://www.ualberta.ca/ORLANDO/), the Electronic Text Corpus of Sumerian Literature (http://www-etcsl.orient.ox.ac.uk/project/sgml-xml.htm), and others have taken this approach. However, the implications for a project's funding, staffing, and training as well as the time-scale for deliverables must be taken into account.

Definition Box:

Definitions

DTD: Document Type Definition, the formal set of rules that define the elements that may occur within an

encoded document and their structural relationships (their relative order and nesting)

Content model: A component of a DTD, giving the structural definition of a particular element

Occurrence indicators: Within a content model, the occurrence indicators show how often a given element may appear (once only, at least once, or any number of times) and whether it is required or optional.

#PCDATA: Parsed Character Data, i.e. words and spaces but not tags.

Tag: an individual piece of encoding that marks the start or end of a textual feature, set off from the document's content by special characters (in practice, usually angle brackets: <tag>).

Element: A textual feature within an encoded document, including the start-tag, the end-tag, and the encoded content: <name>John Smith</name>

Attribute: a modifier to an element, almost as an adjective or an adverb modifies a noun. Attributes come in many varieties and may be used to indicate the type of element (for instance, <name type="person">), its location in a series, a link to a related element, the language of the element's content, an alternate reading, and a wide variety of other kinds of information.

Entity reference: a special character sequence which is used as a placeholder for some other text. Entity references are often used to encode characters which cannot be typed directly in ASCII, such as accented characters, ornaments, or non-roman alphabets. They may also be used as a placeholder for boilerplate text. In SGML and XML, entity references typically begin with an ampersand and end with a semicolon, e.g. é Entity references can also be used to point to external files, such as page images, that can be referenced in the markup and displayed as if they were embedded in the text.

The example below shows a simple DTD fragment that describes a very basic encoding for poetry. Each permitted element is declared together with a specification of what it may contain, and in what order. The occurrence indicators (question mark and plus sign) indicate whether the element in question is optional, and how many times it may occur. The commas separating the elements indicate that the elements must occur in this order. Thus the first line of this DTD specifies that a poem may start with an optional heading, followed by at least one or more <lg> elements, and ending with an optional closer. The second line indicates that the <poem> element also has a required type attribute, which provides a way of identifying the kind of poem more specifically. In this case, the DTD defines a list (unrealistically brief) of possible values, although it is also possible to leave the values unspecified.

Most of the other elements in this DTD are defined so as to contain simply #PCDATA, or Parsed Character Data (in other words, any valid character). However, the <lg> (line group) element has a slightly more complex content model. It may contain <l> or <lg> elements; the vertical bar indicates that either one may occur. The plus sign means that one or more of the group of elements it modifies (in this case, <l> and <lg>) must occur. The net result, therefore, is that an <lg> element may contain one or more verse lines, or one or more nested <lg> elements, or a mixture of the two. It may not contain naked characters (without an enclosing element), nor may it be empty.

Example Box:

A Simple XML DTD Fragment

```
<!ELEMENT poem (head?, lg+, closer?) >

<ATTLIST poem type (sonnet | stanzaic | irregular ) #REQUIRED >

<!ELEMENT head (#PCDATA) >

<!ELEMENT lg (l | lg)+ >

<!ELEMENT l (#PCDATA) >
```

```
<!ELEMENT closer (#PCDATA) >
```

The encoded example text that follows represents one of many possible valid documents conforming to this DTD. Equally valid would be a poem consisting of a single line group containing a single line, without a heading or a closer. When designing a DTD, it's equally important to consider the elements you wish to be able to omit, and the elements you wish to require. In other words, you need to decide not only what constitutes the minimum valid document, but also what constitutes the greatest variation you may need to accommodate.

Example Box:

A Simple XML Encoded Document:

```
<?xml version="1.0" encoding="US2" standalone="yes" ?>
<poem type="stanzaic">
<head>The Clod and the Pebble</head>
<lg>
<l>Love seeketh not itself to please, </l>
<l>Nor for itself hath any care, </l>
<l>But for another gives it ease, </l>
<l>And builds a heaven in hell's despair. </l>
</lg>
<lg>
<l>So sung a little clod of clay, </l>
<l>Trodden with the cattle's feet, </l>
<l>But a pebble of the brook</l>
<l>Warbled out these metres meet: </l>
</lg>
<l>Love seeketh only Self to please, </l>
<l>To bind another to its delight, </l>
<l>Joys in another's loss of ease, </l>
<l>And builds a hell in heaven's despite. </l>
</lg>
<closer>William Blake, Songs of Experience</closer>
</poem>
```

The remainder of this chapter will discuss in more detail some text encoding languages of particular relevance

for cultural heritage materials.

HTML

A brief discussion of HTML is warranted here if only because it is so widely used and so familiar. HTML is essentially a formatting and display language for the web, designed as a small set of tags for simple hyperlinked documents, and as such its value as a form of descriptive markup is extremely limited. It lacks the vocabulary necessary to describe many of the basic features of cultural heritage materials—most significantly their metadata, their genres, and their textual structure. In cases where HTML is adequate for describing such materials (because the materials themselves or the representations desired are extremely simple) a simple TEI-based DTD would be nearly as easy to use and much more upwardly mobile. XHTML offers some improvements over HTML. Although it offers no greater descriptive power (since it provides the same tag set as HTML 4.0), it does allow for the enforcement of XML compliance such as being well-formed, allowing validation against DTDs or schemas, and extensibility through the formal definition of new modules. For projects that choose to use some form of HTML, XHTML will at least offer better interoperation and increased delivery options (for instance, to the increasing variety of web-enabled devices such as mobile phones and hand-held computers).

Projects that create SGML-encoded texts still rely heavily on HTML because SGML encoded texts cannot be viewed online by most web browsers, and there is a grave shortage of SGML-aware software at this time. Projects such as the William Blake Archive or the Victorian Women Writers Project have developed tools to convert SGML documents into HTML for viewing on the web. Others, such as the Women Writers Project, use commercial software that performs the translation to HTML dynamically. With the advent of XML, web publication is likely to become much more straightforward, and conversion to HTML as an intermediate will become unnecessary.

Definition Box:

HTML

- HTML is an SGML application; HTML 4.01 is the latest standard
- HTML is designed to instruct a browser how a page should be rendered, although the lack of standards for browsers has led to considerable variability in the actual on-screen appearance
- HTML is poor at representing information about a text's structure
- Certain HTML features are browser-dependent
- HTML can be generated from SGML/XML for display

XHTML

- XHTML is a reformulation of HTML as an XML 1.0 application; XHTML 1.1 is the latest W3C recommendation
- XHTML can operate in HTML 4 conformant browsers
- Because XHTML documents are XML conforming, they are readily viewed, edited, and validated with standard XML tools

TEI (Text Encoding Initiative)

For projects creating full-text resources, the TEI Guidelines[4] are the predominant choice. The Guidelines have been adopted by a large number of projects representing a range of different kinds of text, and have proved highly adaptable to local requirements. Among the projects surveyed, the use of TEI DTDs in encoding texts is one of the clearest cases of the adoption of standards for a particular type of material. Their use indicates the close match between the TEI's goals in creating the guidelines and the goals that text encoding projects had in mind when creating their texts:

- to represent textual features for research
- to have simple, clear and concrete encoding standards
- to provide a common core of shared textual features
- to enable user-defined extensions or constraints

- to conform to existing and emergent standards
- to allow multiple parallel encoding of the same feature
- to provide rigorous definition and efficient processing
- to allow the encoder to define the level of markup detail
- to provide for documentation of the text and its encoding

At the same time, most projects have found either that the full implementation of TEI is unnecessary, or that the benefit did not justify the extra time and intellectual effort required. Many have turned to the TEI Lite DTD, a simplified view of the full TEI DTD. [5] The purpose of TEI Lite—meeting 90% of the needs of 90% of users—seems to be borne out in practice, and TEI Lite has become the common starting point for a large number of text encoding centers and projects, including the Virginia Etext Center and the Michigan Humanities Text Initiative. While an understanding of the full TEI Guidelines is still desirable, not least for deciding what elements can be ignored, the use of TEI Lite is recommended as a starting point for good practice in text encoding. It is always possible to add further layers of detail at a later stage, if your needs change.

The basic structure of a TEI-encoded document is very simple. Every TEI document must begin with a <teiHeader> element, which contains the document metadata. The header may be very simple, but can also accommodate detailed information about the electronic text's publication, source, subject matter, linguistic characteristics, and revision history. Following the header is a <text> element which in turn contains <front>, <body>, and <back>, which in turn contain elements. In addition to accommodating all of the typical features of texts—paragraphs, lists, headings, the various components of poetry and drama, names, dates, quotations, bibliographic citations, and so forth—the TEI Guidelines also provide for more specialized encoding of features such as prosodic structures, morphological analysis, subject keywording, and similar features which are useful for various kinds of scholarly textual research.

EAD (Encoded Archival Description)

Although the main thrust of DTD development has been in the direction of humanities and scholarly texts, several other DTDs have been developed to cater for heritage institutions with different text encoding requirements. The most significant of these for the archival community has been the Encoded Archival Description (EAD) DTD.

The EAD DTD began as a cooperative venture in 1993 at the University of California, Berkeley. It aimed to develop a non-proprietary encoding standard for archival finding aids that would include information beyond what is provided by traditional machine-readable finding aids, such as MARC.

The project chose SGML as the most appropriate encoding language, as its document type definition (DTD) concept makes it ideal for the consistent encoding of similarly-structured documents, the key to successful electronic processing. An analysis of the structural similarities of finding aids helped construct the initial FINDAID DTD. This simplified, improved and expanded access to archival collections by linking catalog records to finding aids, enabling the searching of multiple networked finding aids and keyword access. The release of version 1.0 of the EAD DTD was delayed until 1998 in order to make it compatible with the emerging XML.

EAD documents consist of two major parts. The first part is the <eadheader> element, which contains metadata about the finding aid and its encoded representation. The second part is the <archdesc> element, which contains the information about the archival materials described.

The EAD header was modeled on that of the Text Encoding Initiative (TEI). It consists of four elements (some of which are further sub-divided):

- <eadid> (i.e. EAD identifier): provides a unique identification number or code for the finding aid
- <filedesc> (i.e. file description): contains much of the bibliographic information about the finding aid
- <profiledesc> (i.e. profile description): records the language of the finding aid, who created the encoded version and when
- <revisiondesc> (i.e. revision description): summarizes any revisions made to the EAD document or its encoding.

The uniformly ordered elements in the <eadheader> make searches more predictable. Such searches can filter large numbers of machine-readable finding aids by specific categories such as title, date and repository. The <eadheader> is obligatory, so archivists are forced to include essential information about their finding aids that were not recorded in paper form. The optional <frontmatter> element can be used to create title pages that

follow local preferences.

Because finding aids generally describe material at several different, but related levels of detail, these unfolding, hierarchical levels are represented within the <archdesc> element. The <archdesc> provides for a descriptive overview of the whole unit followed by more detailed views of the parts. The data elements that describe the whole unit are gathered together under a parent element called <did> (descriptive identification). These <did> elements are the key to good description as they facilitate retrieval of a cohesive body for discovery. Once the high (or unit) level of description is complete, the component parts can be described using the Description of Subordinate Components or <dsc> tag, at whatever level of detail is appropriate for the collection and the resources available.

Dublin Core

While not an encoding system in its own right, the Dublin Core deserves a reference here as part of good practice in creating encoded metadata. The Dublin Core Metadata Element Set defines a set of 15 essential metadata components (for instance, author, title, format) which are broadly useful across disciplines and projects for resource discovery and retrieval. These components can be used to add metadata to HTML files (using the <meta> tag) but can also be used in other contexts to create basic metadata for a wide range of digital resources. Dublin Core does not provide for detailed administrative or technical metadata, and as such is largely suited for exposing resources for search and retrieval, rather than for internal resource management and tracking. In addition, since its goal is to be simple and broadly applicable to a wide variety of resources, it does not provide for the kind of highly structured metadata about specific document types that TEI and EAD offer. Although projects using these encoding systems will probably not need to use the Dublin Core, they may find it useful to be aware of it as a possible output format for distributing metadata about their resources.

METS

The Metadata Encoding and Transmission Standard (METS) is an XML-based encoding standard for digital library metadata. It is both powerful and inclusive, and makes provision for encoding structural, descriptive, and administrative metadata. It is designed not to supersede existing metadata systems such as Dublin Core or the TEI Header, but rather to provide a way of referencing them and including them in the METS document. As a result, it is an extremely versatile way of bringing together a wide range of metadata about a given digital object. Through its structural metadata section, it allows you to express the relationships between multiple representations of the digital object (for instance, encoded TEI files, scanned page images, and audio recordings), as well as relationships between multiple parts of a single digital representation (for instance, the sections of an encoded book). Its administrative metadata section supports the encoding of the kinds of information projects require to manage and track digital objects and their delivery: technical information such as file format and creation; rights metadata such as copyright and licensing information; information about the analog source; and information on the provenance and revision history of the digital objects, including any data migration or transformations which have been performed. METS is a very recently developed standard but is well worth watching and using.

Link Box:

Links to Useful Resources for digital representation and markup of text:

- TEI: http://www.tei-c.org
- EAD: http://lcweb.loc.gov/ead
- Dublin Core: http://dublincore.org
- METS: http://www.loc.gov/standards/mets/

---

[1] Arguably, in other domains such as industrial or computer documentation, where text encoding languages are intended to govern the creation of new digital documents rather than the representation of existing ones,

encoding cannot by its nature be interpretive, since the author and encoder operate with the same agency (even if they are not the same person).

[2] Strictly speaking, while SGML is an international standard (ISO 8879), XML is only a recommendation from an international consortium (the World Wide Web Consortium). In practice, this is a distinction that makes little difference.

[3] It is important to note that the greatest impact of these differences is on software design; to the encoder and the end user, the change from XML and SGML is not difficult to make.

[4] The TEI Guidelines for Electronic Text Encoding and Interchange (P3, the third release but the first official publication) were published in 1994. The latest version, P4, adds XML compatibility and was published in March 2002.

[5] The TEI DTD is designed as a set of modules which can be combined in various ways to accommodated many different types of texts. Thus there is no single TEI DTD, but rather a set of DTDs that represent the various combinations of the modules. The TEI Lite DTD is a simple combination of the most widely required TEI modules.

table of contents previous chapter next chapter

valid xhtml 1.1

VI. Capture and Management of Images

Introduction

Images have tremendous power on many levels and in many contexts. Before the advent of the Web, the online world was largely limited to textual data, but with the growth of the Internet, the demand for images is huge and continues to grow. In addition to the compelling visual pleasure they offer—which puts them in high demand in the culture generally—they also carry dense cultural meaning and are being used increasingly for research, teaching, and learning.

As a creator and distributor of digital images you have many factors to think through.

As with other kinds of materials, these include the selection of objects and visual documentation for digitization, the intended purpose of the digitized items or collection(s) and their expected users and uses, as well as the availability of relevant resources. Your approach will be determined by a number of considerations: the nature of your material, the needs of your users, and your available financial resources and human expertise. Overall, you should remember that your objective is to create a high quality digital resource that will remain useful in the long term. This resource should have as its cornerstone an archived digital master that can be considered "use-neutral" and can serve as the parent to many children that are used in print, web, video and other forms of reproduction.

Other sections in this Guide provide information that bears directly on the issues in this section. Section II provides a broad contextual view on overall project planning, including many of the larger decisions that affect image capture. Section III gives guidance on selecting materials, which will play a crucial role in determining what kinds of image capture techniques are appropriate. Section VIII offers detailed information on quality control and assessment, and Section XII discusses user needs evaluation, which can help you decide issues such as the required image quality. With this information as a backdrop, an overview of the steps to be taken will include the following:

- Investigate the formats of the source material (works on paper, original paintings, slides, negatives, transparencies, photoprints, maps, postcards, etc., as this will affect your choices of method and equipment and your need for specialized expertise. It is likely that the project contains multiple types of material that need to be evaluated both separately and as a group when planning the scope of the project, the equipment needed, and the most effective work flow.
- Establish an understanding of the characteristics and properties of your source material (its fragility, size, mobility, access, etc.) as this will also affect method, equipment, costs and expertise.
- Assess the way it conveys information (how important is color, line, detail, text?) and the kinds of metadata that can or must accompany it.
- Decide whether you will digitize in black-and-white or color (a variety of bit-depth decisions need to be made).
- Decide on whether you have the expertise needed in-house, whether you need some training or whether you will outsource the job.
- Decide on file format, size, storage and management.
- Define benchmarks to ensure image capture consistency across the collection.
- Test a sample of images for capture, display, and print output.
- Decide on standards for documentation (metadata) of the digital files.
- Clarify decisions about sustainability of the project and preservation of the digital files.

Finally, we emphasize that this chapter should be read alongside the chapter on Quality Control and Assurance, which is an integral element of the creation process.

Planning for Image Capture

Choice of Source: Original versus Intermediate

During the selection stage, you will have chosen the items to be digitized on the basis of their content or information value as well as for their potential importance to a set of users. At this stage in the planning you should also consider whether to digitize from the original object or from an existing surrogate, for instance if the original is inaccessible or could suffer damage from handling during the process. Examples of derivative formats are photographic prints of original art works, microfilm of newspapers, and slide images of objects.

Frequently, the physical size and condition of the object will determine this question. If an original artwork is oversized, then a flatbed scanner will not work. In addition, some objects should not be placed in contact with glass. For example, fragile work, determined as such by a conservator or preservation specialist, should not be put in contact with a scanner platen or a glass plate holding materials flat on a camera copystand. It may be possible to digitize such objects if a high-end digital camera is in place, with the requisite equipment, such as book cradle and cool lights. Even “cool lights” may emit harmful ultra-violent radiation or destabilize the temperature and humidity in the room, and if this will put the objects at risk you may need to look into the latest generation of electronic flash-based, area array digital cameras. If such equipment is not available, digitizing from a film intermediary may be a possible alternative. With any intermediary there will always be some information loss, so digitizing from the original is advisable when the highest level of quality, per item, is needed, in cases such as preservation replacement. It should be noted that digitizing from the original can significantly raise costs—not only for capture, but also for preservation—which is why intermediaries are used so widely in large-scale, production-level digitization projects.

Intermediaries pose their own problems, including poor production and scratched, faded, or out-of-focus images, all of which affect the quality of the digital product and need to be closely monitored through the quality control and assurance process. However, staff at some projects included in our survey, for example those at the University of Michigan, photograph all the originals and digitize the resulting 4” x 5” transparencies. This approach is not common among the major projects, and most will digitize from the originals. It should be noted, though, that it is better to do a good job of digitizing a well-made intermediary than a merely adequate job of digitizing original images.

Before deciding to use intermediaries, it is worth considering the following factors:

- Condition of the object: is it rare and fragile? will it be damaged by exposure to intense or unfiltered light?
- Physical size of the object: is it too large to be managed on the available equipment?

If the answer to these questions is yes, then you should consider using a film intermediary.

Quality and Resolution

Decisions as to what resolution, bit-depth or method of tone reproduction, and image format to use for capture, storage and delivery can only be made after project staff have completed an assessment of the material, the needs of potential users, the available technology, costs per image and likely financial support.

The key questions will be:

- What information content do you wish to convey and in what format(s) – screen image, print, or both?
- What functionality does the end product need to support? For instance, does it need to accommodate art history research, which might require scrutiny of minute details such as brush strokes at high magnification? Or is it intended for less exacting audiences where download speed is more important?
- What capture standards and equipment will achieve this?

While you could produce a matrix and try to balance all these parameters, experience shows that you will inevitably need to make trade-offs. Although local circumstances such as resource constraints, long-term plans for the material, and other factors will dictate what trade-offs are appropriate, other things being equal it is probably a good rule of thumb to reduce quantity and maintain quality. A higher-quality digital object offers greater flexibility for future use, since you can always derive lower-quality versions from it, whereas if your needs become more demanding, a lower-quality image will be inadequate. This consideration may carry additional weight if you are dealing with rare or perishable materials, where you may not have a second opportunity to digitize in the future.

There is no set resolution for any individual resource or collection. Each project must identify the minimum level

of quality and information density it requires for its digital surrogates. A detailed explanation of sampling rates and other measures of information density can be found in the appendix on digital data capture. Briefly, resolution is usually expressed in dots or pixels per inch (dpi or ppi) and measures the density of sample—the number of information samples taken per unit of area— that is captured by the scanning equipment. Generally speaking, the higher the ppi, the more detail being captured. The sampling depth, or bit-depth, measures how much information is captured in each sample: for instance, how many colors or levels of grayscale. The higher the sampling depth, the more subtle the gradations across the image, and the larger the resulting file size. The sites interviewed for this Guide used a range of sample rates from 300 dots per inch (dpi) to 600 dpi for their master image files. Lower resolution images were often derived from the master file to produce distribution copies with smaller file sizes for ease of download. The most widely adopted format for storing preservation quality digital masters is uncompressed TIFF. Derivatives, in such file formats as JPEG, are used for access and delivery versions of the image. The resolution of these derivatives generally ranges between 72 ppi and 150 ppi. The resolution appropriate for a high-quality image of the source material should be determined by the size of the smallest significant detail that will need to be visible, given the intended use of the image.

Images can be either reflective (e.g. photographs or drawings) or transmissive (e.g. transparencies or negatives). With reflective material, light is bounced off the surface of the image and its intensity measured using light sensitive diodes. In the case of transmissive material, light is instead passed through the original source. There are three constraints on the highest resolution you can use in scanning both reflective and transmissive material: the maximum sampling capabilities of the scanning device; the size of the files you can manage; and the level of information you actually need. An example of this last point involves old postcards, which were often printed on poor quality paper using cheap processes. If scanned at high resolutions you capture the texture of the paper, which can obscure other information in the image.

Remember that before digitizing you still need to answer questions about the format of the material, its size (or size range if in a batch); whether you need color or black and white (see the discussion of this question in the cost-benefit analysis section of Section III: Selecting Materials); what is the smallest significant detail in the image that you wish to capture (fine line drawing, detailed photographs, fine art); and to what range of uses will the digital objects be put?

Benchmarking

The greater the detail that needs to be captured, the higher the resolution required. Many projects use benchmarking to identify the best resolution for capturing a given image. Benchmarking will identify the smallest essential detail by investigating the attributes of the source and will address the issues of whether the objective is to convey the informational content of the original or to convey information about its creation, such as brush strokes or engraving methods.

One way to establish a benchmark is to take a sample of the material, perhaps that with the greatest detail, and scan at a variety of resolutions. Show these scans, using the medium on which they will be chiefly viewed in practice, to a sample of users and staff to identify which sampling rate best meets the needs for the digital object. This process will be most effective if you take into account the kinds of equipment your users will actually be using: will they have access to high-end monitors and fast broadband connectivity? Once you have identified the optimum resolution then you can set the specifications for that collection. An alternative to 'evaluative benchmarking' is to use a Quality Index (QI). The Cornell Digital Quality Index (DQI, http://www.library.cornell.edu/preservation/tutorial/conversion/conversion-04.html), developed from guidelines for the micrographics industry, though currently restricted to "printed text, line art and book illustrations," can be used for bitonal and grayscale scans, by measuring the smallest meaningful area of the image and using a QI to calculate the optimum dpi.

The UK's Higher Education Digitization Service (HEDS) has a Matrix for Potential Cost Factors (http://heds.herts.ac.uk/resources/matrix.html) which can be useful when considering different resolutions. Another good resource to refer to is the California State Library Scanning Standards (http://www.library.ca.gov/assets/acrobat/scandocrev1122.PDF) but you should use such instruments with great care.

The main principle to remember is that, once the resolution has been identified for a particular batch or collection, it should be fixed for those materials for the duration of the project.

Method of tone reproduction

When you undertake the actual scanning of your images, you will need to choose between several methods of tone reproduction. This phrase refers to the number of bits (or “bit depth”) sampled for each pixel, and to the technique used to capture the full range of tones (or density) in the source materials.

Scanners record tonal values in digital images in one of three general ways: black and white, grayscale, and color. In black-and-white image capture, each pixel in the digital image is represented as either black or white (on or off). You can choose the threshold at which a given point will be considered to be black, and you can use halftoning algorithms to create a “screened” pictorial image, but black and white scanning is generally appropriate only for digitizing text and line art. In 8-bit grayscale capture, the tonal values in the original are recorded with a much larger palette that includes not only black and white, but also 254 intermediate shades of gray. Again, the thresholds can be controlled—either manually or automatically—as a first step to ensure that meaningful information in the highlight or shadow areas of images are adequately captured, and to get optimal results for originals with different degrees of contrast and overall darkness. In 24-bit color scanning, the tonal values in the original are reproduced from combinations of red, green, and blue (RGB) with palettes representing up to 16.7 million colors.

Your decision about which method to use for tone reproduction begins with the choice of black-and-white, grayscale, or color scanning as the default minimum for various classes of material. Your user needs will be an important determinant, but you should not assume that there is a simple correlation between increasing information capture and increasing user benefit. If your original is a faded black-and-white photographic print, for example, color scanning might accurately reproduce the effects of aging. However, file sizes for the color images would be three times larger than their grayscale counterparts. If color is not needed to meet the stated viewing, printing, and publication objectives for your project, grayscale may be the better option since it will be faster to download and smaller to store. Relatively speaking, adjusting resolution is a very straightforward process. Adjusting the settings for tone reproduction, however, can become complicated. Assume that each incremental jump might require changes to equipment, to operator skill, or both.

Having selected bit depth, the most important decisions regarding tone reproduction relate to reproducing the densities of tone in the source materials. Good digital masters are characterized by having pixels distributed across the tonal range in the image from black to white. Stated in the negative, tones are not “clipped” in good masters. Following scanning, all subsequent image processing relies upon pixels either being in the region of interest or not. A scanner that automatically chooses to reproduce the middle range of tones in an image might not have enough dynamic range to capture the meaningful highlights and shadows at both sides of the spectrum. Because experts say that much of the secret to good color reproduction is getting the neutrals right, the method used to capture densities should be carefully planned. For reflective materials such as photographic prints, grayscales might be used to set the “aimpoints” for black, white and middle gray in the digital masters. With transmissive materials, however, these decisions might need to be made by the operator on a case-by-case basis, which is one of the reasons that it can take much longer to scan negatives than prints in order to produce pleasing reproductions.

In addition to user needs, you should also consider which method will actually capture the information of your originals best. Grayscale imaging is appropriate for continuous tone images which lack color. Black-and-white imaging might be warranted for line art if cost savings are a concern; for instance, a black-and-white scan of a typescript page might provide adequate legibility. However, many forms of line art (for instance, manuscript pages) contain subtle color variation which might be of importance in distinguishing overlays of pen strokes and identifying the order in which the marks were made. For these purposes, full color scanning would offer significant benefits.

Projects will occasionally benchmark their imaging by establishing a minimum resolution for the collection as a whole. Most often the resolution depends upon the original, the resources available and the use that you intend to make of the material. Sample testing provides a good way to benchmark sampling rate. Setting minimum requirements is generally accepted as the best practice for establishing the capture requirements, but this must be weighed against the resources available.

Projects with very long-term digitization efforts may have to address the changes in standards for creating digital resources. Improvements in scanning technology and the diminishing cost of storage have radically increased the cultural heritage community’s minimum standards for capture quality, and digital objects captured ten years ago may have been scanned at what would now be an almost unacceptably low level. A museum or library with a large number of legacy images captured at lower resolution will need to decide how to handle the transition. One possible strategy would be to digitize new material at the current recommended standard, and create from these new scans a set of lower-resolution derivatives to use with the legacy materials, if internal consistency is a crucial consideration. Over time, it may be possible to rescan the legacy materials and slowly upgrade the entire

collection. Under most circumstances, it would be a waste of time to scan the new materials at an arbitrarily low resolution. The conversion of material from analog to digital will always be done in an environment of advancing technology and improving standards. You should be sure that you strike a balance between the costs, preservation suitability of the format, and the best usable standards. Generally, best practice is to capture at as high a level of overall quality as is appropriate to handling, quality, budget, or use considerations in order to limit the upgrading that might be required otherwise.

The advice box below gives summary guidelines for selecting the mode of tone reproduction.

Practical Advice Box:

Some practical guidelines for selecting the appropriate scanning method

- Use black-and-white (1-bit) scans if the image is black and white and contains no shading.
- Use grayscale (8-bit) for material with shades of gray or continuous tones: black-and-white photographs, half-tone illustrations, typescript, archival materials in black and white where, say, ink density varies and paper tonality appears to be important.
- Use 24-bit color or higher whenever color is meaningfully present in the original.[1]

Some projects scan at higher than 24 bit-depths (32-bit color 4 color channels of 8 bits each captured as red, green, blue, and 8 bits of grayscale or 48-bit color) even though the current generation of applications cannot render this depth at the moment. This is a hedge against future improvements in technology that you might wish to adopt. For originals of wide color gamut—e.g. paintings or natural history specimens—it may be important to select capture devices that allow the storage of raw files of 12, 14 or 16 bits per pixel, if an extremely high degree of color accuracy is a goal. Dynamic range may be checked and the image archived for future output devices that will be supported. Document the specific characteristics of the capture device, e.g.. non-interpolated highest resolution, spectral response captured IT-8 targets with the 50% gray neutralized, etc.

Transmissive material (for instance, slides and negatives) requires special treatment when scanning. The decisions made for resolution and bit depth are the same as for reflective material, but as slides and negatives tend to be smaller in size the resolution must be higher to achieve a digital image of useful size and quality. Although film scanners have generally been recommended to capture transmissive material, today many mid- to high-end flatbed scanners outperform production slide scanners. Projects at the Library of Congress capture 35mm slides up to 3500 dpi (the range is from 1200 dpi to 3500 dpi). Individual projects will identify the resolution required depending on the original source, the detail of content and the users' requirements.

You also need to ensure that there is a match between the dynamic range of the material that you are scanning and the abilities of your equipment to capture that range. Dynamic range is expressed as a scale from 0.0 (perfect white) to 4.0 (perfect black). Some color scanners miss subtle differences between dark and light colors and as a result do not create faithful digital representations of the content. The better quality the scanner, the greater the dynamic range. Drum scanners can capture between 3.0 and 3.8 and give good color quality; however they are very expensive and impractical for large projects. The scanner's ability to capture dynamic range is an important factor when deciding which equipment to purchase and this is especially true if you are working with transmissive material. Do not believe the manufacturer's claims of dynamic range. Instead, make your own tests by scanning a standard 22 step Kodak film gray scale and subtracting the highest from lowest perceptible densities. A scanner might have a high dynamic range but also introduce unacceptable noise in the dark areas of the image. If you cannot do your own testing, you may be able to find another institution that has recorded results and is willing to share them.

Knowing your sources will allow you to ensure that you have configured your equipment appropriately. For instance, halftone images are made of regular patterns of dots or lines (the word "halftone" is derived from the printing technology method of representing photographs or artwork). They can be either color or grayscale. During digitization, the scanner dot sample combined with the halftone dots can produce undesirable patterns known as the 'moiré effect'. One way of avoiding this is to scan at high resolution, but you may also need image enhancement processing, either post process or at scan time. Most good quality capture software packages will enable descreening to minimize this kind of scanning artifacts, although you may have to do this post-processing, using an image processing tool such as PhotoShop.

Process and Management Issues

Choosing a file format

In choosing the appropriate scanning method you must not only identify the appropriate digitization standard but also the best file format. You should select file formats that are platform-independent and supported by a wide variety of applications, both for browsing and editing. It is generally understood that for master images, uncompressed TIFF (Tagged Image File Format) provides the most suitable preservation format. 'Uncompressed' means that all the information encoded during the scanning process is retained in full. Digital projects store their master archived files as TIFF and then create smaller or derivative files for display and access for the users. These derivatives may be created using a number of different formats.

The file formats typically used for display and access are JPEG (which will gradually be displaced by JPEG2000), GIF, MrSid, PNG, and PDF. These file formats use compression algorithms to reduce the size of the digital files. Some algorithms, such as the LZW (Lempel-Zif-Welch) encoding scheme used by GIF do this to some degree without disposing of information altogether. LZW, for example, uses a lossless algorithm to compress files by up to 33% without throwing away data irrevocably. When the file is decompressed it is possible to reconstruct the compressed data. Other approaches to compression, including fractal and wavelet compression, take more radical approaches and sacrifice data to minimize file sizes. JPEG (Joint Photographics Experts Group) takes advantage of 'lossy' compression algorithms, and in particular fractal compression.MrSid (Multi-Resolution Seamless Image Database) uses wavelet-based image compression, which is especially well-suited for the distribution of very large images. The Library of Congress uses MrSid to deliver maps from its collections. As well as having impressive compression capabilities with limited visible information loss, it stores multiple resolutions of images in a single file and allows viewers to select the resolution (in pixels) that they feel will be most appropriate. There are other file formats that are suitable for delivery of images in different environments and you should consider which of these will be most appropriate for your project. The table below provides an indication of the most commonly used formats.

Definition Box:

There is a range of non-proprietary and proprietary image file formats available. This table includes some of the more common formats for raster images.

Extension	Meaning	Description	Strengths/weaknesses
.tiff, .tif	TIFF (Tagged Image File Format)	Uncompressed file. Originally developed for desktop publishing. 1 to 64 bit depth. Used mostly for high quality imaging and archival storage.	Generally non-compressed, high quality. Large file sizes. Most TIFF readers only read a maximum of 24-bit color. Delivery over web is hampered by file sizes. Although LZW compression can reduce these file sizes by 33% it should not be used for archival masters.
.gif	GIF (Graphics Interchange Format)	This 8-bit file format has support for LZW compression, interlacing and transparency.	Lossless compression. Popular delivery format on web. .png was defined to replace GIF.
.jpg, .jpeg	JPEG (Joint Photographic Experts Group)	Compressed images. 8-24 bit. Variable amount of compression to vary quality and file size.	Lossy compression. Widely used delivery format. Flexible.

MrSid	Multiresolution Seamless Image Database	image-compression technology	Lossy compression. can compress pictures at higher ratios than JPEG; stores multiple resolutions of images in a single file and allows the viewer to select the resolution.
.pcd	ImagePac, PhotoCD	Lossy compression. 24 bit depth. Has 5 layered image resolutions.	Used mainly for delivery of high quality images on CD.
.png	PNG (Portable Network Graphics)	Lossless compression. 24 bit. Replaced GIF due to copyright issues on the LZW compression. Supports interlacing, transparency, gamma.	Some programs cannot read it.
.pdf	PDF (Portable Document Format)	4-64 bit depth. Uncompressed. Used mainly to image documents for delivery.	Need plug-in or adobe application to view.
.pct	PICT	Compressed. Mac standard. Up to 32 bit. (CMYK not used at 32 bit.)	Supported by Macs and a highly limited number of PC applications.

File sizes

You will need to determine how much storage space your image files will require and this is best done by working out the file sizes first. There are different approaches for scanners and linear scanning digital cameras. In the Calculating Box, Estimating Approximate File Sizes, we provide you with an example of the method. The quality of the master images will reflect the work processes you put in place, the equipment you use, and the resolution (dpi) and bit-depth (BD) you select. External variables that may limit this are the amount of storage available and the time for scanning each document.

Calculation Box:

Estimating Approximate File Sizes (in bytes):

Approximate file sizes for material created with a flat-bed scanner can be determined using the following formula:

$FS = (SH \times SW \times BD \times dpi^2)/8$

FS = file size

SH = Source Height (inches)

SW = Source Width (inches)

BD = bit depth

dpi = dots per inch

/8 because 8 bits = 1 byte, the unit in which file sizes are measured.

Large file sizes make compression essential for delivery to users. Derivative files in JPEG format are interoperable within the community and accessible through common access systems, i.e. the Internet. In some cases a much smaller thumbnail version of the image is used for quicker access (for instance, in displaying search results), allowing the user to download a larger file once they have found the image for which they are looking.

Calibration

Calibration is a crucial aspect of the quality control process in any digitization project. You will need to calibrate the work environment, the monitor, the capture devices, and the light sources. Frequently during the life of the project you will find it necessary to recalibrate this equipment as well.

First, establish the physical environment where the digitization work will be carried out. Avoid fluorescent lighting, natural lighting, reflections, and bright colors on the walls. Ideally, darken the space with black out curtains or dark neutral walls, to eliminate ambient lighting, so that the only light source is the scanner bulb or the cold lights used for the digital camera. Where possible, maintain a dust-free environment with operators wearing neutral colored clothing and protective shoe wear. Avoiding carpets on the floor reduces dust accumulation. While it may not be possible to meet all these conditions, it is crucial to ensure the absence of ambient lighting — do not set the equipment up in a sunny room with fluorescent lighting! Consult conservators to insure that the temperature and humidity conditions will be stable and appropriate for the original work.

Once you have established the physical space, calibrate the monitor of the workstation at the beginning of every day, to ensure that the contrast, brightness and gamma settings are consistent. As monitors improve, gamma settings should remain constant, but do not assume that this is the case. A simple calibration program found in image processing software such as PhotoShop can help to do this calibration. Establish the settings for contrast and brightness at the beginning of the project within the adapted space and make sure all operators adhere to them. A set of guidelines with the appropriate settings will ensure the consistency of the digital images. Data about the settings of the capture device should appear as part of the metadata associated with the digital file. For more critical work you should use a spectrophotometer and calibration software such as the Eye-One Monitor (see http://www.gretagmacbeth.com).

Calibrate the equipment used for the digital process regularly. Most flatbed and film scanners can only be calibrated by their manufacturer because calibration is done at hardware rather than software levels. High-end digital cameras may have to be calibrated as the focal length and lighting may change with daily use. We recommend that calibration settings be defined at the start of each batch. Whether the work is being done in-house or outsourced, an appropriate policy or contract arrangements should be in place to ensure equipment is regularly recalibrated.

Color management

A Color Management System (CMS) should be employed by any institution wishing to accurately reproduce color from the source material (original object, artifact or film master) throughout the entire chain of digital hardware devices including monitors, dye-sub and ink-jet printers, film recorders and printing presses. A CMS is a group of software tools and hardware measurement instruments that work together to map the wide gamut of the original color space into the narrower gamut of display and paper output so that a reasonable and perceptible consistency and quality is maintained.

These systems are complex, evolving, and beyond the scope of this Guide. However, there is a growing need, both in terms of economics and image quality, for institutions to develop expertise and deploy a CMS. While lithographers, pre-press experts, image capture specialists and web designers all may use some form of a CMS they often are grounded in different nomenclatures. Therefore a standard form of communication is required to achieve the desired results in image distribution.

Fruitfully, there is a growing body of literature on the subject. The lingua franca is currently found in Adobe

Photoshop version 7, which leverages Apple's Colorsync and Windows ICM operating system core color technologies to provide cross platform color management. These systems use "ICC profiles" that are based on rules made by the International Color Consortium for storing color information.

The most accessible writing on the subject has been by Bruce Fraser who is preparing a mass market reference on color management for autumn 2002 release.[2]

Link Box:

Color Management Web References:

International Color Consortium (ICC): http://www.color.org/

Apple Corporation's ColorSync: http://www.apple.com/colorsync/

X-Rite: http://www.xrite.com

Munsell Color Science Laboratory: http://www.cis.rit.edu/mcsl/

Electronics for Imaging: http://www.efi.com/

GretagMacbeth: http://www.gretagmacbeth.com/

Targets

Targets provide a mechanism for benchmarking the capture process. In simple terms, a target is a sample with known characteristics that can be used to establish a baseline or standard by which to assess capture. Your project should adopt a policy on what targets and scales to use, when to use them, and how their quality will be controlled. Two types of targets are commonly used: edge and detail (resolution) targets and color or grayscale charts. See Resolution Targets Example Box in Section VIII: Quality Control and Assurance.

Ensure that the targets used are appropriate to the material being scanned, as they are different for transparencies, prints, and certain kinds of reflective material. You should use new targets at the start of your project as they lose their accuracy with age. Scan the appropriate target at least once a day, at the start of a new batch of material, or if the settings of the equipment are changed for any reason. Although some think that targets should be attached to every image, this is not possible for non-reflective formats and will have resource implications as images then have to be cropped for display (though it is sometimes possible to automate this process). Including a scale with each digital image can be of value. If it is not in each image it may be sufficient to include it in the first image of a batch and whenever the settings of the equipment are adjusted (e.g. the position of the camera).

Metadata

Metadata associated with digital objects is the key to their sustainability. Metadata assists in the management of the digital data, documents its origin and specifications, and assists with discovery and retrieval. For detailed information on metadata, see the Appendix on Metadata. Briefly, from among the many types of metadata there are three that are generally used with digital images:

- Descriptive metadata describes and identifies information resources, and is used at the local level to enable searching and retrieval and at the web level to enable users to discover resources. Examples of descriptive metadata include Library of Congress Subject Headings, Categories for the Description of Works of Art (CDWA) (http://www.getty.edu/research/institute/standards/cdwa/), the Art and Architecture Thesaurus (AAT) (http://www.getty.edu/research/tools/vocabulary/aat/), and the Dublin Core metadata set. The title and creator of the object are examples of descriptive metadata.
- Structural metadata captures information about the structure of a digital object and the relationships between its parts, for instance between page images and text transcriptions. It facilitates navigation and presentation. Examples of structural metadata are included in the METS standard (http://www.loc.gov/standards/mets/) and SMIL (Synchronized Multimedia Integration Language)

(http://www.w3.org/TR/REC-smil/).
- Administrative metadata facilitates short and long term management and processing, records data on creation and quality control, rights management and preservation. See Cornell University's "Web Hub for Developing Administrative Metadata for Electronic Resource Management" (http://www.library.cornell.edu/cts/elicensestudy/home.html). Another very useful resource is the NISO IMG draft standard on metadata requirements for digital still images: http://www.niso.org/standards/dsftu.html and http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

A number of issues arise when designing a metadata model for a project and these must be addressed at the outset if you are to ensure consistency:

- Static and evolving elements: will the metadata item need to be changed frequently, or is it constant for the whole project?
- Manual and automatic creation: many systems have automatic metadata capture (for technical elements) but some metadata may have to be created by the operator during the capture process. You will need to assess the importance of this manually created metadata and its role in your collection and delivery plans. Keeping manual metadata creation to a minimum will help reduce costs. But for some projects, detailed metadata created by experts will enhance the project's value so greatly that the cost will be more than justified.
- Who captures the metadata? The operator or an administrator? Very often this task is split between two or more people, particular if subject expertise is required.

Projects, on the whole, use a mixture of automatic capture and manual capture by the operator, subject specialist, or cataloger. The level and depth of metadata captured vary from project to project according to resource levels, including both staffing and equipment.

The operator of the digital camera or scanner will usually capture those technical metadata (as part of the administrative metadata set) that are not automatically captured by the scanning software. Descriptive metadata, particularly those relating to the original object, should be prepared in advance by a staff member with specialist knowledge of the collection.

File naming conventions

Before beginning a project you should define a file naming protocol so that your filenames will be consistent and intelligible. The filename can be part of the metadata and the information in the filename can be used to reflect the collection details, e.g. collection name, book title, or page number. One of the most common pieces of metadata in file names is an ID number. In file names you should always ensure that the three letters after the period are reserved for the file type information (.tif, .jpg) as this is important metadata for the operating system to know which applications can handle this file type.

Images of Text

Although the capturing and encoding of text have been dealt with in the previous section, here we will raise some of the issues in capturing pages of text as images, rather than as encoded text.

As already indicated, text can be captured in three ways: as an image only; as fully transcribed text; or as fully transcribed text with an associated image. There are currently no viable technologies for performing text searches on images, so if searchable text is required, some sort of transcription will be a necessity. However, for many purposes the original presentation of the text is also important, in which case an image of the text may be a useful addition. A source that is visual in nature as well as textual, for instance, an artist's book composed of both image and text, is best represented as both data types.

Example Box:

Projects that link images and text include:

- MOA (Making of America), using MOA2/METS: http://moa.umdl.umich.edu/

- Electronic Text Center at the University of Virginia Library: http://etext.lib.virginia.edu/
- American Memory at the Library of Congress: http://memory.loc.gov/
- Online Archive of California: http://www.oac.cdlib.org/
- AMICO: http://www.amico.org/

Projects that have only captured images of text because of the condition of the original material (for example, manuscripts or very early printed books), often store the best quality images so that when OCR technology improves sufficiently, or when more resources become available, they can create machine-readable text from these images.

Although proprietary file formats are generally not recommended for digitization projects, especially for archival versions, the PDF (Portable Document Format) can be very useful for a deliverable page image. You will have to buy the software to create PDF files (Adobe Acrobat), but the software to view PDF files is provided free (Adobe Acrobat Reader). The ability to create PDF files from any application and PDF's platform-independence make it particularly useful for delivering text material over the web.

Definition Box:

Portable Document Format (PDF)

- A proprietary format created by Adobe, but licensed to other software developers.
- Combines a precise page image (including fonts, graphics, and links) with a searchable full-text version of the document.
- Content can also be saved in RTF (Rich Text Format) for re-purposing.
- Users can read PDF files using the free Adobe Acrobat Reader, but creating PDF files requires the purchase of Adobe Acrobat.
- Provides only limited retrieval and representation of document structure.

Post-creation

Quality control and assurance is an integral element of the creation process, and the best digital surrogates should be created to minimize the work involved (see Section VIII on Quality Control below). Standards, technology, staffing and financial resources are all part of this process. Of all of these factors, perhaps the most significant is operator skill and training; investment in both of these areas will pay off significantly in the long run. However, there are also post processing techniques that may be needed to enhance the digital image. This is a practice carried out by a number of projects and generally in batch mode. The standard program used is Adobe PhotoShop© that has various filters and tools to assist in post image processing. Tasks commonly performed after the capture stage include unsharpen mask, gamma correction, noise reduction, deskewing, and cropping. Although you can do a lot in an image manipulation package, it is no substitute for care and attention to best practice at capture time. It is more efficient to set the guidelines for capture and make sure the operators adhere to these than to run image processing to redress errors made in the capture process.

The dynamic range (or density range) of a scanner is a secondary, but important consideration. If a scanner has a high dynamic range, it can sense a wide range of light levels. Dynamic range is usually higher in drum scanners and slide/film scanners. The dynamic range does not actually indicate how many different levels can be resolved but determines the smoothness of transitions between adjacent tones.

How scanner and camera manufactures implement this technology varies. A table in the Appendix on Equipment indicates the types and uses of scanning equipment available.

Conclusion

The successful digitization of still images begins with a careful examination of a set of complex issues. From selection of materials to the best means of capture and storage, these issues must be carefully considered for

each project. Questions the project manager asks may include the following. Which materials will be selected for scanning—originals or derivatives? If fragile originals are digitized, what safeguards should be in place to protect them? What file formats are most appropriate for the material? What are appropriate file sizes and what will be the impact on storage requirements? What is the best way to develop a file-naming protocol? What about sustainability? This section has answered many of these questions or provided a framework within which to begin to seek answers. For further insight, the reader should consult Section VIII on Quality Control and Assurance. That section further clarifies procedures that should yield a quality product.

---

[1] That said, it is worth bearing in mind that Manfred Thaller's work at Duderstadt (http://www.archive.geschichte.mpg.de/duderstadt/dud-e.htm) and work commissioned by the Library of Congress (http://www.loc.gov/) concluded that for manuscripts except under special circumstances grayscale scans provided sufficient information.

[2] Bruce Fraser, Fred Bunting, and Chris Murphy, Real World Color Management (Peachpit Press: forthcoming, autumn 2002). 600pp; ISBN: 0201773406.

[3] For digital cameras this excluded consumer digital cameras that tend to use Contact Image Sensors (CIS) for weight and cost considerations.

[4] This description relates to standard (Legal sized, 8½ x 14") flatbed scanners. Outsize (or large format) flatbed scanners are also available and employ similar technology. Sheet feed scanners employ an optional attachment to a standard scanner to enable automatic batch scanning of loose-leaf material.

[5] Engineering Scanners are also referred to as Wide Format Scanners.

table of contents previous chapter next chapter

valid xhtml 1.0 strict
abp~03/03

VII. Audio/Video Capture and Management

Introduction

Despite differences in format and standards, the fundamental issues for capture, storage and management of audio and video are quite similar and will therefore be considered together. The interviews conducted with audio and video digitization projects highlighted two broad issues.

The first is concerned with terminology and definition: when considering audio and moving image material, what exactly is meant by digitization? For example, the copying of fragile, nitrate-based filmstock to a digital tape format such as digibeta format after restoration, or the audio transfer of a 78 rpm or wax cylinder onto DAT (Digital Audio Tape) is, strictly speaking, a digitization activity. In the context of this *Guide* digitization implies the analog-to-digital conversion of audio and video materials and their encoding in digital audio and video file formats that can be stored, manipulated and delivered using a variety of software and media (e.g. CDs and the Web). In this digital form, audio and video materials can be used and distributed in a richer variety of ways than in their analog form.

Secondly, the usual factors that come into play when considering capture standards and storage or delivery options for any type of digital object (such as the nature and conditions of the originals, the purpose of digitization, the intended mode of access, the needs and expectations of the intended audience) are the same for audio and video digitization projects. However, the fact that high-quality storage is still difficult and expensive, as well as the very high commercial stakes involved in networked delivery of high-quality audio and video to home users, all put a rather different slant on these familiar issues.

One further consideration of time-based media that does not apply to still image media is the act of editing for content. For instance, in a still image of an art artifact, one would normally not apply much "content-editing" to the master file, perhaps only cropping out the color-bar to create a sub-master for automated creation of publicly deliverable derivatives. However, editing a piece of audio or video for content, cutting out pauses or entire sections of the video that will not be used, is a necessary prerequisite for creating any kind of deliverable user version, and represents a large investment of time which needs to be taken into account when creating masters. It is important to keep one master copy of the “raw source” material that has not been edited for content, as a permanent reference file. But it can also be very useful to keep one version of the file as a sub-master that has received basic content editing, but is not technically different from the master file. Saving this sub-master file will save you the work of redoing this basic content editing for each derivative, and allow for quicker, automated generation of edited derivatives. Saving this edited sub-master will have some associated costs in storage and management, but will save the even more expensive cost of re-editing later on.

This section looks at the motives for digitization and why sample projects have chosen particular digitization strategies, briefly describes the decision points and principles of audio and video capture, outlines the main standards and formats and their suitability for different access scenarios, and considers options for delivery and management.

Why digitize?

In common with other types of digitized media, digitized audio and video data should be easy to handle and manipulate and more convenient to deliver than their analog counterparts. By digitizing analog materials, we unlock the content from a fragile storage and delivery format, and make it possible for the content to be copied without loss. Digitization facilitates research and study, allowing quick comparison, searching and editing within the digital object. In their digital form, audio and video content can be more effectively accessed by users than has been possible with analog collections. Once data is in digital form it can be converted more easily to another digital format without loss of quality, unlike all analog formats, which degrade with each use and lose quality when copied (an extreme example of this is audio wax cylinders which have very limited playing life, but every

playing even of a vinyl record contributes to its destruction).

Digitization is used, for example in the National Library of Norway, to preserve fragile and vulnerable materials (e.g., volatile nitrate-based film stock) or materials which need special handling or obsolete playback devices. The challenge here is to produce a high quality digital version. It is very time-consuming to quality check digital audio against the original analog material.

The downsides are financial (e.g., considerable investment in equipment, and large storage is necessary if high-quality masters are to be stored), technical (e.g., methods of compression are still evolving, high-bandwidth networks are not yet universally in place), the difficulty of data recovery from digital tapes in comparison with analog formats, and the continuing uncertainty about the suitability of digital formats for preservation. In digitizing video the cost of matching the quality to that of the original remains a formidable challenge. This is hard enough with tape/video sources, and is still very expensive with film sources. The Library of Congress still specifies analog audio tapes as preservation copies; the Survivors of the SHOAH Visual History Foundation specifies digital Betacam tape copies as the main preservation medium for film. The National Library of Norway argues that digital video formats are not yet good enough, and storage system resources are insufficient in size to make feasible the extensive reformatting of analog material into digital form. Of course, the main problem is that it is very expensive to create a digital version of analog film or video material of comparable quality, though the price of creating accurate digital copies of video, especially VHS, is currently much less than achieving the relative accuracy in copying film. It is common practice among film archives, such as the British Film Institute (www.bfi.org.uk), to create analog copies, known as sub-masters, of their tape and film masters for viewing and exhibition purposes. A digitized version of the tape is just another, and better, way of making a viewing/exhibition copy for all the reasons outlined above.

Access to analog playback devices

Institutions may find themselves with a rich array of materials in analog form, but without the devices to play this material back. Unlike textual and still image material (with the exception of slides and born digital), audio and moving image material require a playback device in addition to a digital capture device. For example, a flatbed scanner can digitize directly a wide range of reflective media of different formats and sizes (e.g., photographs, letters, printed matter, bus tickets). No similar general-purpose capture device for audio and moving image material exists. A collection that included 78 rpm records, compact cassettes, 8mm film and VHS video cassettes would require a playback device for each of these and each would then need to be connected to an appropriate digital capture device. For audio and moving image material that is already in a digital format (such as CD or Digibeta), playback equipment is less of a problem. Although many—frequently incompatible—proprietary digital formats exist, their recent development means suitable playback equipment is still on the market and relatively easy to source. Therefore this section concentrates on identifying analog audio and moving image formats, their properties and the source device required.

Three methods can be used to progress from moving image film to digital. Film can be transferred onto videotape for digitization via a transfer box or multiplexer. Both these options depend upon the material being projected in some way. Transfer boxes project the image into a box containing a mirror and onto a rear image projection screen with the video camera mounted on the other side of the box. The resulting video is subsequently digitized. These transfer boxes are not expensive, but do not in general produce as high a quality material because they produce generational loss in quality.

A better solution is to use a multiplexer. In this device the projector and camera are mounted on a single table. The image is projected by a set of lens and mirrors, directly into the camera without the need for a projection screen. This has advantages for image clarity. In both processes quality suffers because it introduces an extra production generation into the reformatting of the analog material. An alternative to these methods is the use of 8, 16 and 35mm film for a chain film scanner to digitize directly from the analog film material. These machines scan the films and digitize at the scanner, passing the digital signal to the computer. (They work slightly differently for digital video. In this instance they grab individual lines of video to construct a frame and produce broadcast quality digital video.) In 2001 the costs of these machines remains high at between $500,000 and $1,000,000. One of the strengths of chain scanning is that, because the analog to digital conversion is done at the camera rather than on the computer, there is less opportunity for noise to be added by the process to the analog signal. Whereas small institutions can probably set up a transfer box or multiplexer system, even wealthy institutions would find outsourcing to a facilities house to be the only practical option if they wished to go directly from the analog film to the digital material.

Determining a film's original frame rate is also difficult without viewing the film with a projector, particularly for old 8 and 16mm films. The widespread availability of VHS and S-VHS video players makes the playback of these video formats for digitization relatively simple. The rapid adoption of digital formats in broadcasting, post-production and amateur markets is making the availability of even quite recent analog video devices scarce.

As there are fewer analog audio formats, these provide less of a problem than moving images. Compact cassette players, 33 and 45 rpm record players are still widely available new. Even record players with a 78 rpm speed can still be purchased new. The other formats present a greater challenge. If digitizing the sound as played on period equipment is important, the tone arms of phonographs and gramophones can be customized to provide an appropriate feed. Alternatively, the sound can be recorded via an external microphone onto a more convenient intermediary format. Reel to reel tape, wire recorders and cartridges pose similar problems of transfer. By modifying the equipment, it may be possible to provide a direct sound output. Alternatively, the sound can again be captured via an external microphone to an appropriate intermediate format. Here is where a great deal of specialist advice can be helpful. Just as we noted that it is easier to train a good photographer in digitization than it is to train a digital expert in photographic principles and practice, you will find that sound engineers bring to the digital environment strengths that are difficult to replicate.

In the case of all audio and moving image material, whether it is in analog or digital form, projects should carefully consider the advantages of outsourcing digitization. In general audio and moving image digitization require more and more expensive and specialized equipment than is necessary for still image material.

Audio Media	**Properties**	**Source Device**
Wax or Celluloid Cylinders	1890s & 1900s, up to 5”diameter, 2-4 mins. playing time	Phonograph. See http://www.tinfoil.com for details of digital transfer.
Wire	Magnetic coated wire drums or reels. Invented 1898. Widely used by the US military in WWII. Eclipsed by magnetic tape by the mid 1950s.	Wire Recorder
78 rpm shellac resin discs	1898 to late 1950s, 10”(25cm) and 12”(30cm) most common sizes	Gramophone (wind-up) or Hi-Fi. Gramophone’s steel needles need replacing after each side or record played. Hi-Fi needs a 78 rpm turntable and a cartridge with a 78 rpm stylus. For best results on modern equipment a phono pre-amplifier is required to correctly equalize the different types of record.
45 rpm and 33 rpm vinyl discs	7” (20cm) single and 12” long play (30cm). Long play (LPs) introduced in 1948, stereo recordings in 1958.	Hi-Fi. Hi-Fi requires turntable with 45 and 33 rpm speeds.
Reel to Reel magnetic tape	½” to ¼” magnetic tape. BASF and AEG developed 6.5mm ferric tape and Magnetophone player in Germany from 1935. Post-war development in USA by Ampex and 3M. Stereo capability from 1949.	Reel to Reel player for appropriate width of tape.

Compact Cassette	Magnetic polyester tape introduced by Philips in 1963.	Hi-Fi. Hi-Fi requires compact cassette player.
Cartridge	¼" magnetic tape. Fidelipac (4-track, devised 1956, released 1962) and Lear (8-track, 1965) cartridge systems.	Despite similarities 4 and 8 track cartridges are not compatible and require separate players. Predominantly used for in-car audio. 4 track unpopular outside of California and Florida.

Decision points for audio and video capture

There is a series of decisions to make in digitizing audio and video materials, having to do with hardware and software components, sampling rate and precision. To understand these decisions clearly, it may help to first explain the principles of analog and digital recording, and the digitization process itself.

In analog audio recording, a plucked string (for example) vibrates the air around it. These airwaves in turn vibrate a small membrane in a microphone and the membrane translates those vibrations into fluctuating electronic voltages. During recording to tape, these voltages charge magnetic particles on the tape, which when played back will duplicate the original voltages, and hence the original sound. Recording moving images works similarly, except that instead of air vibrating a membrane, fluctuating light strikes an electronic receptor that changes those fluctuations into voltages.

Sound pressure waveforms and other analog signals vary continuously; they change from instant to instant, and as they change between two values, they go through all the values in between. Analog recordings represent real world sounds and images that have been translated into continually changing electronic voltages. Digital recording converts the analog wave into a stream of numbers and records the numbers instead of the wave. The conversion to digital is achieved using a device called an analog-to-digital converter (ADC). To play back the music, the stream of numbers is converted back to an analog wave by a digital-to-analog converter (DAC). The result is a recording with very high fidelity (very high similarity between the original signal and the reproduced signal) and perfect reproduction (the recording sounds the same every single time you play it no matter how many times you play it).

When a sound wave is sampled using an analog-to-digital converter, two variables must be controlled. The first is the sampling rate. This rate controls how many samples of sound are taken per second. The second is the sampling precision. This precision controls how many different gradations (quantization levels) are possible when taking the sample. The sampling error or quantization error means the fidelity of the reproduced wave is not as accurate as the analog original, basically the difference between the analog signal and the closet sample value is known as quantization error. This error is reduced, by increasing both the sampling rate and the precision. As the sampling rate and quantization levels increase, so does perceived sound quality.

In digital representation, the same varying voltages are sampled or measured at a specific rate, (e.g. 48,000 times a second or 48 kHz). The sample value is a number equal to the signal amplitude at the sampling instant. The frequency response of the digital audio file is exactly half the sampling rate (Nyquist Theorem). Because of sampling, a digital signal is segmented into steps that define the overall frequency response of the signal. A signal sampled at 48 kHz has a wider frequency response than one sampled at 44.1 kHz. These samples are represented by bits (0's and 1's) which can be processed and recorded. The more bits a sample contains, the better the picture or sound quality (e.g. 10-bit is better than 8-bit). A good digital signal will have a high number of samples (e.g. a high sampling rate) and a high number of bits (quantizing). Digital to digital processing is lossless and produces perfect copies or clones, because it is the bits that are copied rather than the analog voltages. High bit-depth is also result in much-increased dynamic range and lower quantization noise.

Ideally, each sampled amplitude value must exactly equal the true signal amplitude at the sampling instant. ADCs do not achieve this level of perfection. Normally, a fixed number of bits (binary digits) is used to represent a sample value. Therefore, the infinite set of values possible in the analog signal is not available for the samples. In fact, if there are R bits in each sample, exactly 2R sample values are possible. For high-fidelity applications, such as archival copies of analog recordings, 24 bits per sample, or a so-called 24-bit resolution, should be used. The difference between the analog signal and the closest sample value is known as quantization error. Since it can be regarded as noise added to an otherwise perfect sample value, it is also often called quantization noise. 24-bit digital audio has negligible amounts of quantization noise.

With this background established, we can return to the practical questions a digitization project must ask. The first decision to make in digitizing audio materials involves hardware and software components. Digital audio can be created either by recording directly to the digital sound card or by using an external device to transfer audio material. High quality external devices produce superior results to sound cards; for archival digitization purposes, a high quality stand-alone ADC is recommended. Most internal PCI audio cards are built from inferior quality components and are prone to electrostatic interference from the computer circuitry.

Sample values over times are most commonly encoded in the PCM (pulse code modulation) format. This is the foundation of the digital audio file. PCM data can then be transmitted via a number of digital interfaces (such as AES/EBU) to other devices or software applications.

The next important decision to be made when making an analog to digital audio or video transfer, for example from vinyl or audio cassette or VHS videotape, is on the sampling and bit rates — in other words, the quality of resolution at which the transfer is to be made. Different sampling rates have an important effect on the end result. Nothing can compensate for a bad decision at the sampling stage, so the decision has to be informed, including careful consideration of purpose, intended longevity, circumstances and needs. Put plainly, the higher the number of samples, the better the resulting quality. Current technology allows audio digitization at the so-called “DVD standard” (96,000 Hz/24 bit) and should be recommended as the preferred audio digitization standard for most organizations. However, the quality of the original also needs to be taken into account: there is no point in using a high sampling rate for a poor quality original.

Related to the decision on sampling rate is the purpose of the digital transfer and the intended target audience and mode of delivery (e.g., is a preservation master at the highest possible quality necessary? Are users to access the materials via slow home Internet connections?). Of course deciding at what rate to sample has time, labor, and cost implications. Will it be possible, and cost-effective, to re-digitize the original source material at a later date for another purpose? Will the analog material be accessible in the future? Are they so fragile that you only have one opportunity to digitize from them? Are the access devices becoming increasingly rare? If not, then a better quality initial digitization is recommended to ensure cost-effective future uses. As we have noted elsewhere in this *Guide*, once material has been reformatted it is rare that the work will be done again. It is thus usually better practice to capture at the highest rate you can afford, and deliver a downsampled version, than to capture at a low rate now simply because your immediate intention is to provide the material as small video images over modems on the web.

A policy decision has to be made on whether to clean up or enhance recordings and this, again, depends on the purpose of the digitization: is the objective to restore the recording, to re-create the sounds and images that reached the original recording device, or to make an accurate re-recording of the original? Filtering and noise reduction techniques that remove audio hiss, clicks and pops in an old recording inevitably change that recording and cut out some of the original sound.

Different organizations take different views according to their objectives. The Library of Congress, for example, takes a conservative stance on noise suppression for the preservation masters of historical recordings, seeking to reproduce the original as a recording before cleaning up or enhancing copies for specific listening or study purposes later on. Similarly, for much folk audio material it is important to provide a faithful digital representation of the original archival material, and even if, for example, there is a dog barking in the background of the singer’s performance, it should be left in. From this master version it would be possible to produce derivatives in which the barking dog was removed if you wished to provide one group of listeners with access just to the singer’s performance for easy listening and to produce a copy of the master with all the details of context in terms of environment and capture device (e.g. lack of noise suppression) for folk historians and anthropologists.

Given the flexibility of the digital audio file, it is recommended to digitize at the highest available settings (e.g., 96 kHz/24 bit) without any outboard or software digital signal processing (DSP) applied. The only exception may be a high-quality adjustable compressor/limiter to help with really noisy and soft signals. All other DSP techniques can be easily applied in the post-production process, and their choice should be determined by the delivery purpose, mode, and the target audience.

There may be exceptions to this general rule. For example, if making a digital transfer from an audio cassette it may be appropriate to use Dolby to get the best possible re-recording from the original. Indiana University’s Hoagy Carmichael collection equalizes 78 rpm recordings but not reel-to-reel tapes. Uses of such techniques vary according to needs and circumstances; professional studio practices and methods for adjustment and preparation of equipment (e.g. cleaning and demagnetizing tape heads, distortion filtering, alignment testing) may be beyond the resources of one digitization project but vital for another. Once again this should never be done to the master, but may be done to derivatives depending on the purpose you are trying to achieve with

them.

One should not forget about calibration and adjusting equalization (EQ) curves. Some analog recordings will require the use of calibration and an appropriate EQ curve (e.g., many vinyl recordings) to approximate the signal characteristics intended by the original mastering engineer.

The choice of digitization standards should not be contingent upon the type of acoustic signal to be digitized. While it is true that speech does not have the same dynamic range as the sound of a symphony orchestra, this should not justify the use of a lower bit-depth for speech recordings. We should apply the same, high standards to all kinds of acoustic signals.

Standards and formats: audio

The use of standards increases the portability of digital information across hardware platforms, space, and time. There are in general two types of standards in the marketplace, those that are proprietary and those that are non-proprietary. Proprietary standards are frequently developed by a single company or consortium and are designed to provide that organization or group with market advantages. Non-proprietary ones may also be developed by commercial consortia or not-for-profit groups, but the architecture of the standard is publicly accessible and often in the public domain. Three main audio formats are in common use:

- Microsoft's WAVE (.wav) format has been for a time the *de facto* standard for high-quality audio capture and preservation masters on PCs, and has largely overtaken AIFF (.aif) format. Projects in a Macintosh environment are still using the AIFF format. Of the projects interviewed, Harvard University Library uses the AIFF format for capture and preservation of sound files as AIFF includes structural metadata in the form of channel definitions and time marks.
- MPEG 1 Layer 2, fast being superseded by MPEG 1, Layer 3 (MP3) format offers high quality sound that comes close to .wav quality but at greatly reduced file sizes, achieved by bit-rate "lossy" compression. MP3 files therefore download to users much more quickly than .wav files and compression reduces space needed for storage and network bandwidth.
- Streaming formats such as RealAudio (.ra) which allow listening "on the fly", as the sound reaches the user's computer, eliminating the need to download a complete file.

Definition Box:

Audio Formats:	Extension	Meaning	Description	Strengths/weaknesses
	Liquid Audio Secure Download	Liquid Audio is an audio player and has it's own proprietary encoder. Similar to MP3 it compresses file for ease of delivery over the Internet. Only AAC CD encoder available.	Boasts CD quality. Compressed file, thus some loss.	
.aif, .aifc	Audio Interchange File Format	Developed by Apple, for storing high quality music. Non-compressed format. Cannot be streamed. Can usually be played without additional plug-ins. Allows specification of	.aifc is the same as aif except it has compressed samples.	High quality. Flexible format. Large file sizes.

		sampling rates and sizes.		
.au, .snd	SUN Audio	Mostly found on Unix computers. Specifies an arbitrary sampling rate. Can contain 8, 16, 24 & 32 bit.	In comparison to other 8 bit samples it has a larger dynamic range. Slow decompression rates	
.mp3	MPEG-1 Layer -3	Compressed format. File files vary depending on sampling and bit rate. Can be streamed, but not recommended as it isn't the best format for this — RealAudio and Windows media are better.	Typical compression of 10:1. Samples at 32000, 44100 and 48000 Hz.	Small file sizes. Good quality.
.paf	PARIS (Professional Audio Recording Integrated System)	Used with the Ensoniq PARIS digital audio editing system. Can contain 8, 16 & 24 bit.		
.ra	Real Audio	One of the most common formats especially for web distribution. Compresses up to 10:1.	Sound quality is passable, but not high quality. Lossy compression.	
.sdii	Sound Designer II	Originally digital sampling and editing platform. The format is still in use. Used mostly on Macs by professionals. It's a widely accepted standard for transferring audio files between editing software.	Problems with playing on PCs. High quality. Large file sizes.	
.sf	IRCAM	Usually used by academic users. 8 or 16 bit, specifies an arbitrary sampling rate.		
	Older format, .wav files are far more common. Used			

.voc	mostly in IBM machines. It samples in relation to an internal clock.	Is not a flexible format.		
.wav	Wave	Windows media non-compressed format. Can usually be played without additional plug-ins. Specifies an arbitrary sampling rate. 8, 16, & 32 bit.	High quality. Large file sizes. Can be used on both Macs and PCs	
MIDI	Musical Instrument Digital Interface	Good for instrumental music. The file play digitally stored samples of instruments which are located on a sound card.		

It may be useful to be able to make a simple comparison between the file sizes of three of the formats. For example, a five minute music file will be some 60MB if stored in .wav, 5MB as an MP3 file, and about 1MB as a RealAudio file.

The MPEG standards are among the most important for digital audio and video. The Moving Picture Experts Group (MPEG, http://www.cselt.it/mpeg/) develops standards for digital audio and video compression under the auspices of the International Organization for Standardization (ISO). Each of the MPEG standards is designed for a particular purpose and is continually being developed. It is most commonly encountered as a means of delivering compressed video over the World Wide Web but these standards have also made interactive video on CD-ROM and Digital Television possible. The commonly encountered audio format MP3 is in fact a version of the MPEG-1 audio layer 3 standard.

MPEG standards have progressed considerably and care needs to be taken when using the term “MPEG format” (see table *Development of MPEG Standards*). MPEG 1, 2 and 4 are standard formats for encoding audio-visual media, MPEG 7 is a metadata standard for describing audio-visual media while MPEG 21 is a descriptive framework to encompass the creation, delivery, use, generation and transactions of digital objects. Projects that are intending to encode audio-visual material should be aware that MPEG 1,2 and 4 essentially define the decompression standard: the technology at the user’s end that puts the compressed stream of data back together. It is individual companies that control the encoding technology that compresses the data to be sent. When MPEG 1 was introduced, technology companies such as Microsoft and Apple envisaged a utopian future and included decoders in their software. When MPEG 2 was introduced the likes of Microsoft, Apple and Real Networks decided the cost of MPEG 2 decoding licenses was too high and enhanced their existing technology. These provide high-quality, but proprietary AV streams supported by the distribution of free players (decoders for users). These systems can encode MPEG 2 but distributing MPEG 2 encoded files is problematic because it has been overtaken by proprietary formats such as Real. Therefore, for most projects seeking to encode AV material in an MPEG format, it is MPEG 1 that is a realistic option.

Definition Box:

The Development of MPEG Standards:

MPEG Format	Properties

MPEG 1: Started in 1988 and released in 1992. A standard for the storage and retrieval of moving pictures and associated audio on storage media	Designed for coding progressive video at a transmission rate of about 1.5 million bits per second. It was designed specifically for Video-CD and CD-I media. MPEG-1 audio layer-3 (MP3) has also evolved from early MPEG work.
MPEG 2. Started in 1990 and released in 1994. A standard for digital television.	Designed for coding interlaced images at transmission rates above 4 million bits per second. MPEG-2 is used for digital TV broadcast and digital versatile disk (DVD). An MPEG-2 player can handle MPEG-1 data as well.
MPEG 3. Merged with MPEG 2 in 1992.	A proposed MPEG-3 standard, intended for High Definition TV (HDTV), was merged with the MPEG-2 standard when it became apparent that the MPEG-2 standard met the HDTV requirements.
MPEG 4. Started in 1993, with version 1 released in 1998 and version 2 in 1999. A standard for multimedia applications that is currently being extended.	Designed to meet the convergence of telecommunications, computer and TV/Film industries and provide for flexible representation of audio-visual material.
MPEG 7. Started in 1997 and parts 1-6 (out of 7) released in 2001. A metadata standard for describing multimedia content data.	Designed to support some degree of interpretation of multimedia content's meaning by a device or computer by as wide a range of applications as possible.
MPEG 21. Started in 2000. A framework that is capable of supporting the delivery and use of all content types across the entire multimedia development chain.	Designed to provide a framework for the all-electronic creation, production, delivery and trade of content. Within the framework the other MPEG standards can be used where appropriate.

Sampling rates

As noted above, ideal sampling and bit rates depend on the nature of the original, but they are increasing as the technology allows. Simply put, sampling rate refers to the interval between points at which data are collected and bit-depth to the number of samples taken at any one sampling point. The comparison between digital audio and digital imaging is probably obvious; audio sampling rate (say 44.1kHz) is analogous to the number of pixels per inch (ppi) captured from a digital image (say 300 ppi) and in both cases the bit-depth relates to the number of samples taken at each interval point (say 16-bit stereo for audio or 24-bit color for images). Until recently a standard high-quality sampling rate was the CD-quality equivalent: 44.1kHz, 16-bit stereo; indeed this is the quality at which the Variations Project at the Indiana University (Bloomington) uses for capture and preservation. However, 48 kHz 16-bit is the sampling rate routinely used by the National Library of Norway for old recordings such as wax cylinders, and where the originals are of better quality 24-bit is used.

Very limited quality audio originals such as those in the Library of Congress's Edison collection were created from DAT tape at 22 kHz, 16-bit, mono. However, depending on the characteristics of the source item, the Library of Congress specifies 96 or 48 kHz as a sampling frequency for a master file as a future replacement for reel-to-reel analog tape recordings currently designated as preservation masters. In 2001, the American Folklife Center at the Library of Congress hosted a national meeting to discuss best practices for audio digitization. The consensus of the meeting was to move to 96/24 for preservation/archival purposes. We see no reason, given the declining cost of storage space, not to recommend 96/24 as best practice. Harvard University Library uses a sampling rate of 88.2 kHz for capturing and preserving and a bit rate of 24 for capturing and preserving and 16 for delivering. The file sizes created at these sampling rates are approximately 30 MB per minute at capturing stage and 1MB per minute at delivery. Research shows a clear difference in moving from 16 to 24 bit depth.

Standards and formats: video

Definition Box:

Moving Image Media[1]	Properties	Source Device
8mm & Super 8 Film	18 fps (frames per second) most common frame rate, followed by 12 fps and 24 fps (often used with sound film). The latter frame rate tended to be used by professionals or for capturing moving objects. During the early 1960s 18 fps started to appear. 8mm sound film appeared around 1960. Super 8 introduced by Kodak in 1965. It is perforated in the center and not the edges of the frame. 3” diameter reels are most common, 6” and 7” reels and cassettes are also found.	An 8mm film projector, for "standard" 8mm and/or Super 8 film. Determining the original frame rate can be problematic. Most older projectors are variable speed which is useful. Projectors should be in excellent condition and the film unshrunken. Capstan drives are better for the film and the sprockets.
16mm Film	Very common film format	16mm film projector
35mm Film	Very common film format	35mm film projector
¼” Reel to Reel Video Tape	Can be confused with audio tape. 10” reels are audio, some video, as well as audio, formats used 7” and 5” reels.	¼” videotape recorder.
½” (12.5mm) Reel-to-Reel Video Tape		½” videotape recorder. Machine maintenance and replacement parts very difficult.
¾” (U-Matic) Tape or Cassette	Broadcast TV format. U-Matic has been around since the early 1970s and remains a popular production and archive format because of relatively low cost compared to Betacam SP.	¾” U-Matic machine. Come in fixed or portable, reel or cassette versions
1” Reel to Reel Video Tape		1” Reel to Reel tape player.
2” Reel to Reel Video Tape	Television programs from the late 1950s to 1970s.	2” Reel to Reel tape player. Playback equipment for this has become increasingly rare.
8mm Video	8mm video comes in two formats 8mm and Hi-8 (equivalent to VHS and S-VHS)	Hi-8 players can play standard 8mm cassettes but not vice

Cassette		versa.
½" (12.5mm) Video Tape Cassette	Betacam SP is a popular field and post-production format. M-II is a popular broadcast quality format. S-VHS is a higher quality format of the ubiquitous VHS home video cassette. The now obsolete Beta and Video 2000 formats also used ½" tape cassettes.	Betacam SP and M-II require compatible players. S-VHS players will play standard VHS cassettes but not vice versa. Although almost identical, S-VHS cassettes have additional holes in the casing.[2]

Three main file formats are in common use: MPEG (see table), QuickTime and RealVideo. However, both the Library of Congress and the National Library of Norway have held back from keeping preservation copies of film material as files on servers, but rather have kept digital video preservation master copies on Digibeta tape. The Library of Congress uses the sampling ratio of 4:2:2 for digital tape copies, which is the current component digital tape recording standard. 4:2:2 refers to the sampling ratio of the three parts of a component color difference signal (one luminance channel and two chroma channels). For every 4 samples of the luminance channel there are 2 samples for each of the chroma channels. As usual, as the sampling rate increases, so the quality increases. In 4:4:4, the chroma channels are sampled equally to the luminance channel, creating better color definition, but this high sampling rate cannot easily be recorded onto tape.

Of the file formats that projects might use for service, rather than preservation copies, the highest quality films are likely to be stored in the .mpg (MPEG) format. The average file size for the MPEG 1 format is about 9 MB for each minute of film. The Library of Congress's MPEG 1 files are created at 30 frames per second at a data rate of approximately 1.2 Mbits per second of playing time. The National Library of Norway makes digital transfer from film copies in MPEG 1 at 1.5 mbits per second, at a resolution of 25 frames per second, or MPEG2 at from 6 to 15 mbit per second.

QuickTime may include a variety of compression methods; some higher end, some lower end. For instance, QuickTime (with Cinepak compression) offers smaller, downloadable files and allows films to be viewed on lower-end computer systems. The Library of Congress's QuickTime files are created at 10-15 frames per second at a data rate of approximately 640 Kbits per second, usually quoted as 80 Kbytes/sec of playing time. The average file size in the QuickTime (Cinepak) format is about 5 MB for each minute of motion picture. The Berkeley Art Museum/Pacific Film Archive (BAM/PFA) currently captures video content as DV (see DV discussion). This DV stream is then converted and saved as a video master file in QuickTime/DV format. Derivative files are extracted at a much smaller resolution and are delivered online in QuickTime/Sorensen format for video, and QuickTime/Qualcomm for audio-only materials. Content digitized so far by BAM/PFA includes videos of artist talks and works of video-art from the collection. Works on film will require a different methodology.

RealVideo is a streaming format allowing viewing of the moving image material as it arrives at the user's computer and thus eliminating the need to download the file completely before viewing. Real Media format is especially useful for computers with slower Internet connections, such as a 28.8kps modem. Video playback is slower (3-6 frames per second), may be affected by Internet traffic, and currently provides an image of lesser quality than the worst broadcast TV. But it does make the distribution of material to wide audiences possible.

Definition Box:

Moving Image formats

Extension	Meaning	Description	Strengths/weaknesses
.mpg	Moving Picture	Standards created by the group working for ISO/IEC. MPEG-1: for Video CD and MP3 are based on this early standard. MPEG-2: DVD based on this. MPEG-4: Standard for	Good quality and low file sizes. MPEG-1 can take

	Experts Group	multimedia on the web. MPEG-7: Currently under development; for 'Multimedia Content Description Interface'.	a while to load.
.qt, .mov	QuickTime	Created initially for Macs, can now be used on PCs too. QuickTime player. Quick Time 4 has streaming capabilities.	Excellent quality, easy capture, widely used, can be large. In Windows the QuickTime player takes up lots of space.
.viv	Vivo	No updates since 1997. Played on VivoActive player. Video stream always sent over http (unlike Real Video or Windows Media). Bought by Real networks in 1998.	High compression rates, poor quality due to compression to maximize streaming, various incompatibility issues.
.avi	Audio/Video Interleave	QuickView, Windows' Media Player. Replaced largely by MPEG and Windows media.	Large files, very good quality, must be encoded/decoded properly,
.rma	RealMedia	Streaming format. Proprietary format that is an equivalent to Windows Media.	Requires RealMedia plug-in.
.wma	Windows Media Video	Streaming format. Version 8 offers near DVD performance.	

Not all audio and video material will need to be digitized from analog material; much of it will come, increasingly, from digital materials. In discussing still image material we noted that derivatives of digitized images should be measured in pixel dimensions rather than dots per inch so digital video formats are measured in pixels. Digital video in NTSC format consists of 720 x 480 pixels[3]. This is the standard resolution used in MPEG-2-compressed commercially distributed DVD movies. As you examine the chart on digital video formats, it will be obvious that the main differences between the DV formats relate to the formats of the tapes (e.g. size and running time) themselves, but there are, in the case of DVCPRO 50 and Digital-S, some differences in the compression algorithm used.

Definition Box:

Digital Video Formats:

Format	Tape Size	Compressor	Compression Ratio	YUV sampling	Running Time (mins.)
DVCAM	6mm	DV25	5:1	4:1:1 NTSC 4:2:0 PAL	184
DVCPRO	6mm	DV25	5:1	4:1:1	183

DVCPRO 50	6mm	DV50	3.1:1	4:2:2	90
Digital S	12.5mm	DV50	3.1:1	4:2:2	124
Digital Betacam	12.5mm	Sony	3:1	4:2:2	94

Post-capture processing

Following the capture of digital AV material, significant post-processing may be necessary or desirable, either to clean up the data using digital processing techniques (such as noise reduction), or to produce smaller, downsampled and compressed versions for various kinds of distribution. Some of these processes can be automated, but they are still labor-intensive and need to be planned and provided for in advance. Examples of post-capture processing carried out at the University of Virginia Robertson Media Center include adding data, fades, and altering frame rate and size using Media Cleaner, Final Cut Pro or Imovie tools. Other projects may choose to clean up noise or edit moving image materials into easily accessible chunks. Post-processing should be performed on digital derivatives, with an unedited digital master copy kept for reference and archival purposes. It is worth considering the kinds of uses to which these derivatives will be put so as to sequence the kinds of processing you perform and minimize the number of special-purpose derivatives you must create and maintain. For instance, where initial capture is performed at a very high bit rate and sampling depth, as recommended in this chapter, some institutions produce two other masters to meet delivery needs: a de-noised "production master," and a downsampled de-noised master from which Real Audio files can be easily made using current software.

Audio-visual metadata

Metadata is a crucial extra element that must be created to accompany digital audio or video, and the principles of metadata interoperability and documentation standards are as important to digital AV media as to still image and text media. However, unlike textual resources, audio and video cannot currently be adequately searched by themselves as a raw resource. There is no cheap, standard way to apply the equivalent of a "full-text search" to AV materials. As a result, metadata for audio and video is doubly crucial to internal management as well as public use of such resources. Metadata for digital audio and visual resources can be used in much the same way as metadata for complex digital objects composed of still images. For instance, a metadata standard like METS (Metadata Encoding and Transmission Standard) can be used to describe the structure of a digital object: for instance, a book that is represented by dozens of digital page images, plus a full-text transcription. METS allows one to connect an individual page image with structural information about that page (i.e. Chapter 10, page 99), as well as alternate representations of that page (say a transcription of the text). In this way, one can present and navigate the various digital files (page images, transcriptions, METS XML file) as one cohesive “digital object”. This allows users of the resource to search for phrases that may occur on a particular page, and be taken to the location of that page in the complex digital object. Similarly, AV metadata standards such as METS (with the appropriate extension schema), or others like SMIL (Synchronized Multimedia Integration Language) can be used to describe the content and structure of time-based digital files such as audio and video. SMIL, for instance, can be used to describe structural metadata about a particular frame of video (frame 30, timecode 01:20:36.01) as well as link the appropriate series of frames to alternate representations such as a transcription of the dialogue in that scene. As with image resources, this allows users to search for a particular bit of dialogue or the name of a character, and be taken directly to the video scene in which they appear. In addition to acting as a discovery tool, audio-visual resources metadata also helps enable the exchange of resources between institutions, and facilitates the internal management and preservation of such resources.

As of 2002, there is no shared good practice for what constitutes minimum metadata for digital audio and video in the cultural heritage sector. So, whether creating A/V metadata for management or for discovery, it is recommended to use or adapt a standards-based metadata standard like METS or SMIL, and to use the standard to capture as much information as the institution can afford. Note that some technical metadata can be automatically—and hence cheaply—generated; for instance, some cameras create and can export metadata about capture date, format, exposure, etc. Metadata about content is more expensive to create, since it requires human intervention. At a minimum, useful retrieval will require a basic metadata set for each file, based on a

standard like the Dublin Core. In addition, any information that will require effort to rediscover later on (for instance, the format of the original material, or its provenance) should be captured immediately if possible. Most institutions will probably decide to create a very basic metadata set for each file, with the possibility of adding further information later in the lifespan of the digital object. Fewer institutions will be able to create a full metadata record for every object from the outset, although arguably this may be the more efficient way to proceed, since it consolidates processes such as quality checking.

Link Box:

Key AV Metadata Sites

- Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
- Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
- Metadata Encoding and Transmission (METS) Standard http://www.loc.gov/mets
- MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
- Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
- Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
- Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
- Library of Congress's Digital Audio-Visual Extensions to METS Standard http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
- Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html

Options for delivery and management

The main issues to be examined when considering options for delivery, storage and management are connected with the larger file sizes associated with audio and moving image material. The restricted nature of available bandwidth for delivery and the lack of appropriate data storage facilities pose challenges for institutions. Rights protection is another important issue, in which streaming as a delivery option can be part of a solution. For example, in Indiana University's Variations project, while the material in copyright is digitized under the legal provision for libraries (preservation or fair use), students may view the streaming format but are unable to copy (and potentially misuse) it. At the Berkeley Art Museum / Pacific Film Archive, copyright issues have slowed efforts to digitize film and video content. Digitization has so far included documentary material and selected works of video art but no film, as of 2001.

While almost all of the project managers interviewed are committed to maintaining top quality archival masters, many are prepared at this time to trade off top quality against lower, compressed quality for ease of access and delivery. Similarly, they are prepared to trade material of longer duration (say full-length feature films) against availability of short (say 30-second) samples. These are purely pragmatic decisions, based on current technology and bandwidth.

Much depends on the purpose of the digitization project: if the objective is to provide ready access to audio visual materials (perhaps historical material), which are otherwise very difficult to view or hear for a general audience, then delivery of lower quality, compressed versions fits the purpose. If, however, the aim is to preserve fragile materials, there is still some doubt as to the durability of most digital formats; most projects are keeping options open by storing preservation copies on tape until such time as server-based digital formats are proven for long-term preservation purposes.

For many projects, storage is problematic because of the massive data sizes of audio and video material. At the extreme ends of the spectrum, the Survivors of the SHOAH Visual History Foundation and Indiana University's Variations project serve as exemplars of large data storage options for video and audio material respectively. At Indiana, migration of MPEG files from the current tape storage system to a new university-wide mass storage system has recently been completed, with migration of WAV files currently in progress. The Digital Library Program is working with University Information Technology Services to ensure long-term preservation/access of

Library objects in this mass storage system.

The system for storage and retrieval at the Survivors of the SHOAH Visual History Foundation involves use of local caches of 1 Terabyte for instant retrieval. If the information is not available on the cache then the disc server (180 Terabytes) at the Visual History Foundation is accessed. The final step is for the information to be found in the tape storage and uploaded to the server and local cache. This is done automatically by a system that uses a robotic arm to locate, pick, and load the desired information, but it can take between 5 and 10 minutes. This system is appropriate for a project with such a massive scope (52,000 video records of testimonies) and multi-million dollar budget, but much of its practice cannot be applied to smaller museums or archive projects as the level of funding required will not be available to produce such sophisticated systems and technology.

In summary, it is difficult to draw general conclusions or to make firm recommendations on ideal storage, delivery and management of digitized audio and moving image materials. Practice and technology are still evolving, although pointers to the future indicate that streamed delivery of highest quality audio and video material will become widespread, problems of massive storage will ease as costs decrease and reliable preservation formats will be found. Perhaps the most prudent course in the meantime is to transfer materials at the best sampling rate possible and store them at the best possible quality with as little use of lossy compression as the budget allows in order to keep the highest quality materials to work with in the future and eventually migrate as necessary.

---

[1] This table does not include details on esoteric formats such as lantern slides, picture discs, 4.75, 9.5, 17.5, 22, 28 and 70mm film formats.

[2] There are three different VHS formats: PAL which is common in Australia, New Zealand, the United Kingdom and most of Europe; NTSC used in the USA, Canada and Japan; and SECAM used in France, much of Eastern Europe and Russia. PAL (Phase Alternate Line) uses 625 horizontal lines at a field rate of 50 fields per second (or 25 frames per second). Only 576 of these lines are used for picture information with the remaining 49 lines used for sync or holding additional information such as closed captioning. SECAM, (Sequential Couleur avec Memoire or sequential color with memory) uses the same bandwidth as PAL but transmits the color information sequentially. NTSC (National Television Standards Committee) is a black-and-white and color compatible 525-line system that scans a nominal 29.97 interlaced television picture frames per second.

[3] DV-PAL has a pixel dimension of 720 x 576.

table of contents previous chapter next chapter

valid xhtml 1.1
abp~03/03

VIII. Quality Control and Assurance

Introduction

This section discusses some of the most effective methods of Quality Control and Assurance (QC&A). In the fields of cultural heritage and humanities the authenticity, integrity, and reliability of digital material is crucial to users at all levels, whether scholars, teachers, students, or researchers. What is sometimes less obvious is that effective QC&A must be planned at the outset of the project and built into all its processes. For QC&A to ensure the quality of the project's deliverables this activity cannot be left until the end. If your QC&A procedures accomplish nothing except to reveal the errors that faulty processes have already produced, they are largely a failure.

At the outset it is useful to make the distinction between quality control (QC) and quality assurance (QA) as these terms are often interchanged. Quality control includes the procedures and practices that you put in place to ensure the consistency, integrity and reliability of the digitization process. Quality assurance refers to the procedures by which you check the quality of the final product.

For all classes of digital materials three measures of quality can be used: completeness, fidelity or faithfulness to the original, and legibility. Completeness simply measures whether the entire object has been captured, without cropping or other essential loss of material. Legibility is a functional measure which indicates whether the digitized version is intelligible: for text, whether the characters can be read; for images, whether the object depicted can be discerned at a basically acceptable level. Fidelity goes a step further and measures whether the digital version represents the original in a way that goes beyond simple legibility: a legible representation of a manuscript page will allow the text to be deciphered, while a faithful representation might also convey the visual texture of the paper and the gradations of ink color. Within these measures one can also make the distinction between subjective and objective measures. Comparing a digital image with the original using the naked eye is a subjective measure. Using a computer to log the number of potential errors per 1000 characters in OCR-generated text is an objective measure. In practice, both subjective and objective measures are combined in the whole quality control and assurance process. A digitizer evaluating color reproduction in a digital image might use a program such as Adobe Photoshop to confirm that the RGB values for a color bar included in the image fall within the correct range. A computer may flag errors in a text but it is a human editor who checks the nature of these errors (if indeed they are errors) and decides whether to accept, reject or correct the text.

Related to this point is the need to be clear about the differences between tasks that are fully automated (e.g., comparing checksums to ensure success of data transfer), some that are semi-automated (correcting what software identifies as likely OCR-generated errors), and finally, those that are fully manual (confirming that text has not been inadvertently cropped; confirming that image content is not skewed in the frame).

Finally, throughout this section it is important to be aware of the differences between the quality assessment of people and systems and the quality assessment of products. In the former case, you are intervening to assess the capacities of your staff and of the systems and products you use during the digitization and quality assurance process. For instance, you might administer a color blindness test to the staff who will be conducting the color correction and quality assurance of digital images, or you might use technical targets to ensure that a camera is operating consistently from batch to batch, or week to week. In the latter case, you are establishing systems to check the actual product that results from the digitization process. Examples of product quality assessment, which abound in the chapter, include using targets as a means to establish correct tone reproduction for the image products created by the photographer and digital camera, or checking a certain percentage of each batch of scanned images for skew and cropping as they come from the vendor.

Digital objects

QC&A in some form is an integral part of any digitization project, but the procedures that projects put in place

are often informal, of variable consistency, unreliable, and under-resourced. The aim should be to embed QC&A in the project at the points where it will do the most good. The project must define which digital products, such as masters and deliverable images, metadata, encoded texts, transcriptions and translations, are to be included in its QC&A program. Once you have determined the range of the QC&A program, the next step is to establish appropriate QC benchmarks.

Although there are no broadly accepted standards for image quality, encoded text accuracy or audio quality, there are a number of generally accepted guidelines that can prove very useful. These will be discussed in more detail below, but in general your QC threshold needs to match the purpose of the digital deliverable. For instance, the QC requirements for digital images that will serve as replacements for deteriorating nitrate film would probably be much more stringent—given that there may never be another digitization opportunity—than for “access” images created from relatively stable black and white photographic prints.

In general your QC threshold needs to match the purpose of the digital deliverable.

The first step is to carry out an initial series of QC benchmarking tests on a sample of the analog materials to test capture settings and establish threshold guidelines for rejecting digital material that does not meet the quality criteria. This should be undertaken for each type of material to be digitized and for each of the different types of output that the deliverable will take. For example, tests should be carried out with text for screen and print, or with audio for streaming or for preservation. In this way, different QC procedures, and possibly different methods of QA, will need to be established for different types of material. The resulting benchmark standards will then form the basis of an ongoing QC&A program. These QC benchmarks must be documented and presented to project staff in a way that makes them easy to implement and easy to monitor for effectiveness.

These QC benchmarks must be documented and presented to project staff in a way that makes them easy to implement and easy to monitor for effectiveness.

As well as defining your QC benchmark, you will have to decide on the scope of your QA program. Do you check the quality of every image or page of text against the original (100% check)? Do you undertake a stratified or random sample (for example, every 10th image, or a random 10%)? What QA procedure are you going to follow when a digital object is rejected? If a digital image is created from a surrogate will you compare the digital image against the original object or the surrogate for the purpose of QC&A? Also remember that even if your digitization has been outsourced, the project is still responsible for performing a QA check on the vendor’s work, irrespective of how detailed the QC&A requirements are in the contract with the vendor. Images of England, for example, has put in place quality assurance procedures and a custom application that allows it to check the digital images, mark errors on screen and return the error details to the vendor to enable them quickly to identify and correct the error. Furthermore, will the digital images be compared against the originals or some other benchmark such as a color chart? At what magnification, if any, will the comparison take place?

It is also necessary to ensure that a project’s QC&A program includes measures for controlling the digitization environment and the way staff work (e.g. ensuring they take regular breaks). If you have specified your hardware correctly (see Equipment in Section II on Resources, and in more detail in the Equipment Appendix) then you should have a system that is appropriate for the types of materials being digitized and the purpose for which they are being created.

In addition to ensuring the correctness and informational quality of the digital materials you are creating, you need to ensure the integrity of the data files themselves over the long term. This is a significant preservation issue, and is discussed further in Section XIV. But it is also an ongoing quality assurance concern and is worth touching on here. There are two points to be addressed: checking the integrity of files as they move through your workflow and are transferred from medium to medium, and checking the storage media at intervals to guard against failure and data loss. The first of these can be automated to a large extent. Checksums, for instance,

provide a simple way of ascertaining whether a file has been altered or corrupted during data transfer. For SGML and XML documents, parsing and validation can help verify the document's integrity, though they cannot detect every kind of error. Another more powerful option is a version control system, which can track all changes made to a set of files and prevent casual or accidental alteration. In a lengthy workflow process, where the files change hands and move around a great deal, basic steps like these can save you huge amounts of effort and trouble later on.

The second point—the quality of the media on which your digital objects are stored—is easily neglected. However, storage media vary greatly in their longevity and robustness, and you should not only assess which media are appropriate for your purposes but also take preventative measures to guard against loss. Although hard disk failure is relatively rare, running periodic scandisk checks and defragmenting drives on a monthly basis can go a long way to identifying bad sectors and preventing read/write errors as well as improving the performance of your computer in the process. Removable media such as CDs and tapes should be purchased from a reputable brand; batch diagnosis for structural validity of the media and files is a relatively efficient method of quality assurance. Certain media, such as floppy disks, JAZ cartridges, and ZIP disks, have relatively high failure rates which make them inappropriate for storage and backup.

There are a number of further steps that can be taken to minimize potential errors in QC&A for particular types of material, including images, OCR and encoded text.

Images

Image capture may serve a range of project goals, and for some of these only completeness and legibility are essential. However, for projects that require that the digital image be a faithful representation of the original (to the extent allowed by the medium), careful calibration of the equipment will be needed to ensure consistent and highly controlled results. You should also ensure that the material that you are to digitize is free of dirt and dust, that it is positioned on a calibrated scanner or camera, that the digital capture environment is adequately controlled (e.g. free from stray light sources), and that suitable control targets have been used (see box).

Example Box:

Resolution and color targets should be used.

Common resolution targets:

- AIIM Scanner Test Chart#2
- RIT Alphanumeric Resolution Test Object
- IEEE Std 167A.1995
- IEEE Standard Facsimile Test Chart

Common color targets:

- Q13 and Q14 Kodak Color Separation Guide and Grayscale Targets
- Q60 Kodak Color Target
- Kodak Grayscale Charts

Remember that color targets are made with organic dyes and that these dyes breakdown as they age. Therefore over time the charts lose their accuracy.

For all images there is a series of key QC&A tests to perform. The first set of checks is relatively straightforward. Check that the entire image has been captured (i.e. not cropped) including any captions or titles. Are pages missing or out of sequence? Is the image skewed? Does the image have the correct file name? The second set of checks is more complex to assess, and includes detail reproduction, tone reproduction, color reproduction, and color accuracy.

For images of textual material, line drawings, etchings, plans and other objects with distinct line-based features, detail reproduction is the key to image quality. When benchmarking, a resolution target or the smallest resolvable

detail should be used. This provides a comparison point for examining legibility, completeness, sharpness, contrast, serifs and uniformity, paying particular attention to individual strokes and dense cross hatchings.

For grayscale and color images the bit depth and dynamic range are as important as resolution in assessing image quality. These issues have already been discussed in some depth in Section VI on images, and in the appendix on digital data capture. Briefly, bit depth is the amount of information (in bits) used to represent a pixel. The use of a grayscale or color chart can provide a standardized reference point for assessing the quality of color and tone reproduction. Assessing color and tone reproduction can be highly subjective, particularly if fidelity is desired, but features to look out for include the presence of details in shadows and highlights (an indication of a good dynamic range), and a smooth transition in tones, particularly on skin and sky (a blotchy or pixellated effect is an indication of insufficient bit-depth). Compare color, contrast and brightness to the original or to a color chart, paying particular attention if digitizing from negatives, where simple inversion can produce a color cast, or digitizing from print, where a herringbone, or moiré, effect can be present.

Example Box:

What to look for when checking digital images for quality:

- image not the correct size
- image in wrong resolution
- image in wrong file format
- image in wrong mode or bit-depth (e.g. bitonal when it should be grayscale)
- overall light problems (e.g. too dark)
- loss of detail in highlights or shadows
- poor contrast
- uneven tone or flares
- missing scan lines or dropped-out pixels
- lack of sharpness
- excessive sharpening (including halos around dark edges)
- image in wrong orientation
- image not centered or skewed
- incomplete or cropped images
- excessive noise (see dark areas)
- misaligned color channels
- image processing and scanner artifacts (e.g. extraneous lines, noise, banding)

One of the best examples of an imaging QC&A system is that of the Genealogical Society of Utah (http://www.lds.org). Quality control and assurance of the images is an integrated part of the GSU capture system and uses software specially developed for the programs. The volunteers examine each image and reject for skew, readability and color balance. If rejected, the image will be recaptured and noted in the log file for re-indexing.

When the images are sent to GSU from the projects, an audit program, again specially developed for the project, carries out further checks. A wizard sets up the rejection threshold and uses a random number generator to identify the individual items to be selected for the inclusion in the statistical sample. An auditor checks the image, and there are twenty-four possible rejection criteria; if an image is rejected then the reason for rejection is noted in the log file. If three images are rejected, the audit program turns off and the whole batch must be re-imaged. Auditors are trained to use this system and to evaluate the images. Typically, they look at 150 images in a batch at their choice of speed, e.g. 1 sec per image. They also use histogram analysis as well as checksum, to facilitate automatic QC&A. (Checksum is a value computed from a block of data and transmitted and stored along with it to check whether errors have occurred in transmission or storage.)

The work environment has a significant impact on the QC&A of digital images.

You can control some QC&A environment factors with the following methods:

- Ensure the work area is free from strong colors and reflections (e.g. a room painted gray without direct sunlight)

- Always calibrate the monitor regularly (see Section VI on Images)
- Use appropriate software for your image or text format
- View originals and print output in bright, preferably natural, light
- View display output in low, but not dark, light
- Minimize the number of different people who perform QC&A measures, to reduce variability
- Ensure that all QC&A staff use identical equipment and configure it the same way
- Consider using color management software to ensure consistency throughout the digitization chain
- Document the QC&A procedure either through having online checklists or through paper-based forms that staff complete for each object

OCR and encoded text

When digitizing text, the page images are subject to the same QC&A checks as for line art images, but further checks are required at the OCR and encoding stages. The range, scope and method of QC&A must be established and appropriate benchmarks set. If you are creating OCR texts for indexing purposes or batch processing texts, then an error rate of 00.5% may be acceptable (this is the Library of Congress' NDLP threshold for batch processing by vendors), but if you are creating a scholarly edition, then nothing less than 100% accuracy may be required. OCR rarely produces results better than 99.9% accuracy, or one error in every 1,000 characters (roughly 10–12 lines). Double processing documents and checking each version against the other can speed up identification of errors, but frequently there is no substitute for manually checking the digital version against the original by second, or even third proofreaders.

The Thesaurus Musicarum Latinarum (TML http://www.music.indiana.edu/tml/) at Indiana University employs one such text QC&A system. The quality control procedures in place for the digital deliverables involve at least three sets of complete checks on each text. The individual entering the data is expected to check and correct the text before printing it out and passing it to the person responsible for proofreading and checking. This second person proofs the printout and marks any errors identified. As the corrections are made to the electronic version the marks on the printout are counter-marked. Then, both the printed text with its marks and counter-marks and the electronic text are passed to a third person for review, prior to approval and addition to the database. Where there is a high error rate at any stage in the process, the text is printed once again and subjected to a second proofreading, as outlined just above. The final check and approval by the project director facilitates consistency in the quality control process. Lessons they have learned from this experience are that it has been difficult to persuade people to proofread character-by-character (rather than word-by-word) and to refrain from global search-and-replace editing. In general, the TML has discovered that no more than four double-spaced pages of 12-point text can be proofread per hour with an acceptable rate of accuracy.

With XML or SGML encoded texts the use of a parser to validate files against a DTD greatly assists the QC&A process. Remember, however, that while the coding can be declared well formed and valid, a parser would not pick up typographical errors in the content of the text, or an incorrect choice of tag from a set of structurally valid choices. To catch the latter kind of error, more sophisticated (and usually project-specific) tools are sometimes developed which allow a reviewer to look at overall patterns of tag usage and spot anomalies without having to review every file in detail. For large-scale encoding projects such tools may repay the cost of development.

Audio and moving image materials

Digitizing audio and video materials in the cultural heritage community is a relatively new area of activity, QA in this area is especially new, and so there is little documentation of QA practices in libraries and museums regarding AV material. The feasibility of these methods for a community that must weigh scarce labor resources against the need to protect the investment in digitization is thus unfortunately not yet well established in practice. However, some significant examples do exist.

The Library of Congress' 'Audio-Visual Prototyping Project' specifies the following QA in a contract document intended for vendors who are handling the digitizing of LOC audio materials:

"Contractor quality review of audio shall include, but is not limited to, the following criteria:

- complete item has been captured
- re-recordings are not flawed by noise or distortion beyond that present in the source recordings
- all files open and play properly

re-recordings meet specifications for resolution, sampling frequency, and other formatting requirements
- recorded calibration tone sets meet requirements for reproduction quality"

Phase 2 of the European project, "Presto: preservation technology for audio and video" (http://presto.joanneum.ac.at/index.asp) includes the goal of automating AV quality control: "Implement automated quality control: audio and image analysis algorithm development, fitting of algorithms to the application, and cost / benefit analysis." Indeed, Presto partners are developing tools that can, for instance, track errors that occur in transferring film to digital and log the timecode of the frame of the film in which the error occurs so that humans may conduct special quality control measures on that area of the file.

While work in this area begins in individual institutions and projects, it is clear that there is no consensus in the cultural heritage sector about what constitutes good practice for quality control and quality assurance of digital audio and video files (metadata is covered elsewhere). Standard QA methods such as random selection of sample files for inspection and playback by a human, as well as other QA methods developed for digital images, are highly recommended as the very minimum while more robust and yet cost-effective means are being tested on a broad scale in the cultural heritage sector.

Metadata

QC&A of metadata is generally less well documented, but its accuracy is perhaps more important than that of the digital deliverables themselves. If users cannot find an object because of poorly checked metadata they will never know that you have meticulously checked and verified the audio, image, or text. There are even fewer guidelines for checking metadata quality than for images and texts, although for XML-encoded metadata the general principles for encoded text apply to a large extent. From among the projects in this *Guide* good practice indicates use of the following:

- An appropriate metadata standard (such as EAD, Dublin Core, TEI Headers, METS)
- Name authority files (for person and place names)
- Subject thesauri and classifications
- Data type controls on fields (e.g. text, number)
- Independent double checking of entries

Good practice projects have checked metadata at the time of image checking, but many projects have very little metadata QC&A. It is recognized that metadata QC&A is not done once, as with images, but is an ongoing process. As with any QC&A, this should be considered in the resources both at the creation of the project and for the future.

Conclusion

The QC&A techniques presented in this section have focused largely on detailed questions of accuracy and the scrutiny of the project data. However, you should be careful not to neglect QC&A in other areas of the project. Steering or advisory groups, project plans, and flow charts can all perform an important role in assuring the quality of project management. Similarly, documented procedures for the archiving and preservation of digital material, such as the frequency of back-ups, rewinding archive tapes, and moving copies off-site, all contribute to the overall QC&A environment of a project. Finally, be sure to take a global view of the entire project and the complete product it is creating: how do the parts work together? Is the overall design successful? Does it meet your users' needs? These larger evaluative questions require different methods from those sketched above, but are an important aspect of the overall assessment of the project's success.

table of contents previous chapter next chapter

valid xhtml 1.1
abp~03/03

IX. Working with Others

Introduction

As digitization moves from small, discrete projects, conducted within individual institutions, to larger, multi-departmental, multi-institutional and international digital programs, collaboration becomes an increasingly vital consideration. The funding, creation, access and sustainability of digital surrogates increasingly depend on arrangements in which institutions work with others. These may range from relatively simple cooperative arrangements to collaborations where partners develop a common goal and input resources jointly. They may also include working with consultants, paid or unpaid, who contribute specialized expertise but are not closely involved with project planning, or using outside agencies to accomplish specific segments of the project. This section of the *Guide* uses the term “collaborate” in the broad sense of any type of cooperative or partnership relationship.

Our interviews revealed that all the projects and programs that are establishing good practice in all areas of digitization already collaborate on various levels. As they improve their understanding of their collections and the digitization process, they increasingly recognize that collaboration has a crucial role to play. It quickly enables a team to widen its subject base, increase its available expertise, and leverage its equipment, human, and financial resources (see Section II on Project Planning).

Collaboration and cooperation have always played an important role in the cultural heritage sector, from international art exhibitions to simple inter-library loans for academic establishments. Fruitful use and understanding of our heritage depends on such collaboration. Digitization extends this rationale, and provides more efficient ways of improving access and increasing scholarship.

Any digital program starting out should explore all the possible layers of collaboration and also understand the advantages and possible difficulties that any collaboration can hold.

Why collaborate?

Institutions seeking to create digital versions of their collections may not consider collaborating as a necessary, or even desirable, option. Their aim may be quite limited and local in its scope: for instance, to digitize a particular section of a collection for a specific research purpose or a specific teaching task, where collaboration seems like an unnecessary complication. However, collaboration does not necessarily entail large-scale partnership over a number of years, but can be as simple as seeking advice and adopting standards that will streamline the current project and enable future collaborations. In these cases, the advantages may lie in the greater breadth of experience applied to the project—an asset under any circumstances—or, more importantly, in ensuring that the local collection can interoperate effectively with related resources elsewhere. Collaboration may also help solve problems of permissions or digital rights management, by grouping together organizations with a common interest and giving them greater leverage when negotiating with rights-holders. For instance, Georgetown University encountered resistance from publishers when seeking permission to digitize critical editions of philosophical works. However, once the university joined with professional organizations such as the Hegel Society of America, they were able to negotiate successfully for digitization rights.

Collaboration does not necessarily entail large-scale partnership over a number of years, but can be as simple as seeking advice and adopting standards that will enable future collaborations.

Funding may also play an important role in encouraging collaboration. Seeking funding is a critical and time-consuming aspect of any digital project or program, from identifying likely funding sources to writing complex bids and complying with funders' guidelines and requirements. Sharing this work may be advantageous, particularly for smaller institutions. Increasingly, the major funders in Europe and in North America are seeking collaborative programs for successful digital projects. The purpose may be to exploit the full capacity of digitization, to unite geographically disparate collections and improve access, as well as to enhance partnership between institutions. It may be that collaboration is a necessary—or at least a highly desirable—aspect of any successfully funded program. The projects we interviewed all agreed that collaboration was absolutely necessary to achieve the depths of digitization that they desire for their collections.

Types of collaboration

There are many kinds of collaboration, ranging from simple consultancy to sharing skills or equipment, outsourcing, and fully-fledged multi-institutional international projects. Collaboration can take place locally, regionally, nationally and internationally. All these forms of collaboration require an understanding of the issues and careful planning and management.

Consultancy

Many projects seek the advice of consultants. One typical approach involves identifying a suitable company or person to visit the project, discuss the tasks involved, and produce a feasibility study or similar document on how the project should proceed. Consultants do not necessarily have to be external; in a large institution, there may be internal experts who can offer advice on planning, creating and sustainability at the beginning of any project. Generally, it is a good idea to seek this internal advice before applying for funding to ensure that the bid is as complete and precise as possible. However, not all institutions have access internally to such expertise and may seek external advice. Remember that external advice will be costly and funders may have reservations about permitting feasibility studies unless they have been specified in the bid. It is worth exploring funders' attitudes to such costs in advance. Advice from an internal source may be negotiated in a more relaxed manner, perhaps with specific advice on funding bids and purchasing technical equipment. If you will need advice regularly from these internal sources throughout the project, then you may have to factor some sort of payment for recovery of salary costs. It is worthwhile exploring your institution to discover what expertise exists. It may also be possible to create a support group of similar projects that can work together, creating an institutional policy that can be applied to new projects, thus saving costs in time and staffing.

Any contract with an external consultant must be well prepared and managed throughout the process to ensure that the deliverables, such as a feasibility study, are produced on time and meet the specifications. This will take more time and cost more than you may anticipate, so think carefully about whether the costs of seeking professional external advice will actually achieve the goals you set. You may find you can reduce costs by using consultants who are located close to you, to reduce costs in travel and phone calls. However, one specialist notes that this could constrain one from getting the right person for the job, particularly if the expertise you seek is rare. A large-scale, multi-institutional project in the U.S. might serve its needs very well by bringing in an expert from the UK, for example, to help plan workflows or to train key staff. It is also advisable to seek references on any institution or company you plan to consult. Staff in universities, libraries and other institutions within the cultural heritage sector will be able to act as consultants or make recommendations. There are other avenues of expertise that projects can explore, such as this *Guide*, or other resources created by the community. Make use of the listservs and newsgroups to seek such advice. The Research Libraries Group (RLG) and NINCH have web pages that will help.

Joint projects — funding, resource sharing from local to international

The most likely collaboration is one that involves more than one project from more than one institution and then generally within the same sector. These projects use the expertise and skills from similar institutions or departments. The Making of America is one such project, involving the University of Michigan and Cornell University. The project describes this ongoing collaboration as follows:

"Drawing on the depth of primary materials at the Michigan and Cornell libraries, these two institutions are developing a thematically-related digital library documenting American social history from the antebellum period

through reconstruction. At the University of Michigan, approximately 1,600 books and ten journals with imprints primarily between 1850 and 1877 were selected, scanned, and made available through the present system. Librarians, researchers, and instructors continue to work together to determine the content of this digital library and to evaluate the impact of this resource on research and teaching at both institutions." (http://moa.umdl.umich.edu/about.html)

As a highly successful project that involved collaboration at all levels and at all stages of the digital process, The Making of America is worth further investigation if your institution is looking for a collaboration model to follow.

Collaboration in different shapes

The impetus for collaboration comes from a number of directions. Funding decisions are frequently made on the basis of the skills and resources available at an institution. Even large institutions may not have all of the skills and expertise required entirely in-house. Smaller institutions may only have their collections and specialist subject knowledge to offer. Even if institutions have sufficient capacity for digital content creation, the capacity required for effective delivery and sustainability may only be found at larger institutions or through pooling resources. Some digital projects will require a critical mass of content to begin with and others will require that their content be scalable in order to be sustainable. As digital projects give way to digital programs, particularly for delivery, the need to collaborate increases. Indeed, underpinning the widespread and convenient delivery of digital content is the need to adopt common standards and practices, which itself drives collaboration. In today's digital landscape the individual, entrepreneurial start-up has given way to the age of partnerships and mergers.

Underpinning the widespread and convenient delivery of digital content is the need to adopt common standards and practices, itself a driver of collaboration.

Digitization projects can collaborate in a wide variety of ways. Collaboration may involve the pooling of resources, or the sharing of content; it may operate at the institutional, local, regional, national or international levels; it may exist within or across sector boundaries; and it may involve formal or informal relationships. None of these types of collaboration in best in the abstract; different types will suit different projects and will affect the partners in different ways. Nevertheless, a number of trends are evident.

The library and museum fields offer fertile opportunities for cross-sector collaboration. The enhanced functionality that web delivery can provide, and users frequently expect, requires an increased amount of interpretative information. Museums and galleries are well versed in providing succinct information on the selected parts of their collections on display while libraries and archives have traditionally concentrated on providing users with access information to their entire holdings. Web delivery demands a combination of these elements: comprehensive descriptive information to facilitate resource discovery, and access and supplementary contextual information to aid interpretation. Cross sector collaboration in physical resources, personnel, skills and experience can thus greatly benefit both parties.

For many smaller institutions some form of content collaboration may be the only method by which their collections can be digitized cost effectively.

The ability of the web to deliver still images, text, moving images, audio and 3D representations offers another opportunity for digital content creators to provide new ways of accessing collections. Traditionally, researchers would access sources across physically dispersed collections that may be distinguished by the type, subject or provenance of the material held. Even if the digital representations of these collections remain distributed (and the distributed collaborative model is a common one) the opportunity to provide a unified point of access, and hence distribution, greatly enhances the functionality for users and the dissemination objectives of collection

holders. Although the educational benefit of access to digital collections is often presumed rather than evaluated it is clear that a single point of access to interrelated and previously unobtainable resources makes the learning experience more efficient if nothing else. The institutional emphasis on education, particularly in the K-12 sector, makes such content collaboration a major consideration.

For many smaller institutions, some form of content collaboration may be the only method by which their collections can be digitized cost-effectively. The collaboration may take the form of licensing rights to a commercial organization in return for creating digital content, or providing access to collections to a larger non-profit institution. Even if projects can afford to undertake the actual digital conversion themselves, issues such as copyright and intellectual property rights, electronic access controls and effective delivery and asset management systems may be beyond their financial and personnel capabilities. Lastly, the ongoing delivery and preservation of digital content raises issues that lie beyond the scope of even the largest institution and require the collaboration of the whole cultural heritage sector if they are to be addressed effectively.

There seems to be no predominant pattern in the level at which collaboration should take place. However, smaller organizations do seem to require either a major institutional partner to act as a catalyst or a partnership with enough similar-sized organizations to gain critical mass. In both these cases a formal collaborative arrangement would be most appropriate. Collaborative groups of any size can bring in expertise less formally by including specialists on their advisory boards.

Outsourcing

Outsourcing is a growing method of creating digital resources. Projects or programs will identify a vendor (usually through a request for proposals or RFP), negotiate a contract and give the vendor access to the material from digitization and to metadata creation, indexing and cataloging. This is a very simplistic view of the process; there are a number of factors that must be taken into consideration before approaching a vendor to ensure that both parties understand the process and agree on the outcomes. Outsourcing can be done on-site as well as off-site; in the former case the vendor will bring equipment and staff to the material. This may or may not be more costly, since the vendor avoids a whole range of infrastructure costs, but may have additional travel costs.

Vendors are increasing their knowledge of the types of documents in this sector and are now purchasing the specialist equipment to handle them. However, you should ensure that any vendor you are approaching understands the nature of your material. Scanning 10,000 business documents, on good quality laser print, is very different from handling 10,000 fourteenth-century church records on different types of material and of varying sizes. Show the vendor a sample of the more difficult material to ensure that they have the requisite expertise and appropriate equipment. You should also ask them to produce sample scans of some of your material to help you assess their abilities.

Make sure that the quality of work produced is of the standard required. This *Guide* has sections on standards for resolution, tonal range, color depth, and other variables (see Sections V, VI and VII). Quality control and assurance (QC&A) is a vital part of any agreement with a vendor (see Section VIII). The actual terms of the agreement will depend on cost and the vendor's capabilities. It is expensive to check every image, and depending upon the nature of material and how you are processing it checking a sample of the digital products may be sufficient, but do not neglect quality control and assurance as a cost saving measure. Agree with the vendor on a guaranteed level of quality control of the process(es) they will provide and require that they have quality assurance procedures in place. The only secure way for you to ensure that the agreed level of quality has been reached is to conduct a check on the output once it has been returned to your organization. For instance, check one in ten images and if any error (such as skew) is found in a specified batch, then that whole batch should be returned to the vendor to be digitized again. For obvious reasons, you should be sure to stipulate such quality assurance conditions in advance.

The most immediate benefit of using a vendor is that once the costs have been agreed, you are protected against unexpected cost increases as these would normally be carried by the vendor. This condition must be stated explicitly in the contract. Establishing a comprehensible and workable contract is an essential stage in the outsourcing process; if it is carefully worked out (using lawyers where possible), then outsourcing can be a great benefit to an institution.

Many projects in the survey reported that they had been reluctant to use outside vendors because of the nature of their material and a lack of confidence in the capability of commercial services to appreciate the conservation, handling, and content issues. Increasingly, though, commercial companies have realized that the cultural heritage sector has funds available for the creation of digital content. Firms with digital imaging experience are

investing in the expertise and facilities that allows the sector to entrust them with the digital process.

Nevertheless, outsourcing should be avoided when it requires rare and fragile material to be sent off-site. In these instances, a vendor should set up digitizing stations on-site so that the materials do not have to be moved. Generally, using vendors is most successful where the material is in good condition and of consistent type and quality. Variations in size and quality will cause the vendor as many problems as they would cause the project team, but vendors rarely have the content and conservation expertise to respond to these difficulties.

Do not underestimate the complexity of getting the analog material from storage to the vendor, having it returned to the institution after scanning, then checking it and returning it to storage. Catalogers and conservators should be involved in this process. The latter will be especially helpful if there are questions as to how material should be packaged, shipped or handled by the vendor. The operation should be planned, managed, and documented to suit the whole collection as well as the material to be digitized.

As increasing the volume of material to be digitized should reduce the per-unit costs, it is always worth asking at the planning stage whether extending the project would be sensible. Collaboration between projects within the same institution will also help reduce per-unit digitization costs by providing access to increased volumes of material.

Outsourcing is thus a viable option but one that requires building a solid relationship with a trusted vendor and managing every step to ensure the highest quality result.

Managing collaboration and partnership agreements

The success of any collaborative venture lies in its management, with firm aims and objectives set out for all partners. Goals and deadlines must be fixed, agreed and adhered to. Employing a project manager to take overall responsibility for ensuring that all partners deliver the work agreed is a distinct advantage. It is also worth ensuring that all partners are doing work that is appropriate to the level of skills and resources available to them. Establishing effective methods and mechanisms of communication between members of the consortium has always proved a challenge for collaborative initiatives. You should work to build good, well-documented channels for communication and decision-making, which involve all of the relevant participants.

All this should be set out clearly in a Partnership Agreement, and the more comprehensive it is, the more likely the project will be a success. The Agreement should make explicit the role(s) and responsibilities of each partner. It should differentiate between those activities that a partner will carry out in support of achieving their local project obligations and those that a partner must complete for the group as a whole (such as local training or dependable file access). The Agreement should be signed by director-level staff to ensure that ‘project buy-in’ has taken place at a senior-enough level in the organization.

Where you are working in a multi-institution collaboration, a consortium policy can make explicit the responsibilities of the central agent. The central agent is the institution that manages the initiative’s financial resources; other sites involved are partners. It is crucial to examine which functions will be centralized and which will be de-centralized and thus under local control. For instance, will training take place locally or will it all be centralized? Will standards of QC&A and metadata creation be defined locally or centrally? Will the texts be stored on a central server for cross-resource searching? Will the digital images be stored centrally, or locally, or both? Some of these questions have obvious answers, but answers to others will vary depending upon the circumstances, material, and nature of the partnership. The Colorado Digitization Project (CDP), for example, has centralized training sessions for all partners and stores the texts and images on a centralized server that all partners can search; some partners also hold the images locally, with the CDP acting as a depository and collections manager. The MOAC project, based in Berkeley, stores all EAD and MOA2 XML documents centrally for searching by all the partners but these records link to images that are stored locally on each museum’s server.

Further examples of the issues that must be addressed include the following:

- If the digital images are stored centrally, can the central agent make use of them for purposes other than those envisaged when the project was started?
- Can the central agent alter the context of the images without express permission from each partner?
- Who creates the upgrades to interface to the central resource?

These questions, and others, should be addressed at the start of any project. This will help forestall any

problems or misunderstandings that might occur later in the process.

The Partnership Agreement applies to the funding and publicity stages of the project just as much as it does to the creation and usage of material. Issues to consider at this stage are:

- How much income will each site have, and who decides what percentage each site should devote to specific activities? For example, it may be that one site spends more on creation and another on interface design, depending upon the strengths of each partner.
- Does the central agent have a responsibility to funnel a certain average percentage of income to the partners for content development? Or will it hold on to funds for system development?
- Are the decentralized partners obliged to provide publicity or credit the central agency whenever publicizing the resource?
- Must the central agent always list the partners whenever publicizing the project?
- Must the partners participate in all fundraising efforts, or just some?

The Colorado Digitization Project (CDP) is an excellent example of how various partners can come together, combining their talents, resources and collections in such a way that all partners are satisfied with the management and organization of the collaboration.

Collaborations such as this enable large-scale digital programs to realize the full potential of digitization. The CDP brings together a variety of institutions from the state of Colorado with the aim of producing a digital resource that encapsulates material from the collections of museums and libraries throughout the state. A member of the CDP staff acts as the project manager, assisting and advising the projects, as well as setting minimum standards for creating images and metadata. Skills and resources are shared through training sessions and scanning laboratories located throughout Colorado. The laboratories provide a critical link in the chain because they house the technical equipment needed to enable the participation of projects from smaller institutions. The project enables both large public libraries with state-of-the-art equipment and small local museums with no equipment at all to collaborate in creating a high quality digital resource. Overall management of the digital resource is controlled from a central point, but this is crucial to the success of any collaborative project; no matter how large or small the contribution of the institution is, the goals and deliverables must be agreed upon to ensure the initiative's success.

A successful collaboration depends on establishing guidelines and creating standards that all partners adhere to. These guidelines will cover such areas as definition of types and depth of metadata, descriptive terminology or control vocabulary, file formats (whether for images, audio or video), standards for interface design, and guidelines for QC&A. This rule applies to all collaborations, whether between two departments in a single institution or as part of a multi-institutional, international project.

It is also important to apply long-term thinking to collaborations and to recognize that it may be beneficial to participate in more than one type of collaboration, so as to maximize each partnership. It is also advisable to think about the life-cycle of a collaboration and the conditions that might bring it to a graceful close. For instance, does the collaboration have an expiration date? If not, is there a date when all agreements are revisited? Has an exit strategy been devised for each partner as well as for the central agency? These questions have to be considered if an institution means to make the best of its collections and resources and ensure that they remain accessible in the future.

Successful collaboration can enable scattered collections to be reunited as well as producing materials that enhance and complement each other, providing rich digital resources for the users.

Thinking Box:

Key factors to consider when starting a project:

- Investigate whether there are areas where collaboration could be possible
- Plan for the possibility of collaboration at the outset
- Look for strategic fit when selecting collaborators
- Commit to equal participation in shaping projects and sharing the benefits
- Involve all partners at all stages
- Establish clear communication lines and document communication
- Moderate and manage expectations, including those of donors and funders
- Adopt shared project goals and focus (for example, create a mission statement)

- Establish a steering group that reflects institutional types, experience, resources, and knowledge of the partners
- Do not amplify institutional hesitation — act, don't just talk
- Have flexibility to accommodate different aims, structures and needs of partners
- Document areas of responsibility, and track progress
- Ensure accountability — one person to enforce milestones and timelines
- Embed interoperability — minimum guidelines for flexibility
- Cement collaboration with face-to-face meetings with minutes and formal agreements
- Pool resources
- Avoid duplication of effort
- Trust, respect, and endeavor to understand professional skills of the other collaborators
- Think ahead about access and distribution to ensure equitability of access and benefits
- Ensure partners have compatible exit strategies

---

table of contents previous chapter next chapter

valid xhtml 1.1

X. Distribution

Introduction

This section looks at how users gain access to your digital objects and how they use them. Once the data (text, images, sound, moving images) and metadata have been created there are two sets of decisions to be made: choosing appropriate delivery option(s) and modes of access. These are key decisions and are likely to be made at an early stage in the development cycle. Delivery options cover questions of how to distribute the data assets and grant users access to them; in other words they are concerned with dissemination, or 'publishing'. Access modes—including applications and interface design decisions—affect how users will be presented with the material and how they may search, retrieve, navigate and manipulate them.

Delivery options

Keep in mind that the delivery options discussed in this section pertain to delivery versions of data and to public versions of metadata. Information pertaining to the storage and distribution possibilities pertaining to ***masters*** is in Sections XIII-XIV. Essentially, there are two choices for distribution: using a network as the medium (principally the Internet); or using portable media (usually magnetic disks, optical discs, and tapes). We consider both choices, including a note of requirements for each, and then compare their advantages and disadvantages.

Portable and removable media

There are three categories of portable and removable media that projects may wish to use to deliver their digital content. These are: tapes (e.g. Digital Audio Tape/Digital Data Storage (DAT), Digital Linear Tape (DLT); optical discs (CD-ROM or DVD); and magnetic disks (floppy diskettes, Zip disks, JAZ drives). There are a number of issues to consider, including the robustness of these media, and in particular their relatively high sensitivity to magnetism, atmospheric conditions and temperature; cost; the level of hardware support among the user base; user preferences; projected lifespan of the digital material; and the widely varying carrying capacities (**for capacity in 2001** see the Table on Commonly Used Types of Removable Media, below). However, for most projects, CD-ROM or DVD are the likeliest options, simply because users are more likely to have access to a CD-ROM drive or DVD drive. In addition, these media are fairly durable, and since they can be distributed in a read-only format, it is harder for users to modify or delete materials accidentally. In order to deliver data on removable media you need either the appropriate equipment for duplicating data onto blank media, or the means to outsource duplication to a vendor. You may also need an installation program (if the applications do not run from CD-ROM). Finally, you need a means of delivering the media to end users and keeping track of orders and fulfillment.

Definition Box:

Commonly Used Types of Removable Media (2001)

Media Name	Full Name	Media Category	Storage type	Current Capacity
DAT/DDS	Digital Audio Tape/	tape	magnetic	For DDS-3 12 GB uncompressed[1] and 24 GB

	Digital Data Storage			compressed
DLT	Digital Linear Tape	tape	magnetic	35 GB - 70 GB
CD-ROM	Compact Disk-Read Only Memory	disk	optical	650 MB
DVD	Digital Versatile Disk	disk	optical	6 GB
Removable Disk Packs		disk	magnetic	10 GB - 70 GB
Floppy Disks		disk	magnetic	1.44 MB
Zip disks[2]		disk	magnetic	100MB and 250MB[3]
Jaz drives		disk	magnetic	1 GB and 2 GB

Networked delivery

Here the Internet or TCP/IP-based intranet are the chief options, although there are other types of network such as AppleTalk or IPX/SPX. There are a number of different delivery mechanisms to users, each suited to different purposes. HTTP[4] protocols are the standard mode for communicating web documents, and given the ubiquity of browsers they represent a near-universal method of disseminating text and images, without requiring special action on the part of the user. Streaming media —for example via RealAudio, RealVideo or MP3—allows you to supply real-time digital audio and video, without necessarily allowing the user to download and save the data locally. FTP[5], which supports file transfer only, is suitable for making raw data, including installable applications, available to users, who may download files and save them locally. In order to deliver data of any of these types via the network, you need the following: a network path between the provider of the data and the users; server hardware; server software (an operating system and server applications); and client applications enabling users to access the data (though as indicated above, this may simply mean free browser technology). If you prefer not to host the delivery yourself, you may be able to outsource it, either to an internet service provider (ISP)[6] or to a partner or consortium member with the appropriate infrastructure.

Pros and cons of portable media

The advantages of portable media for producers of digital assets center chiefly on their cost and reliability, and on the fact that they provide a mechanism to deliver very high quality content to users. The media (e.g. CD-ROMs) are typically fairly cheap. Sale of digital assets on physical media is relatively straightforward, presenting analogies with printed documents which may also ease copyright negotiations, and assets can be licensed multiple times, extending the possibilities of generating revenue. There are also few potential points of failure in the use of removable media, no need to rely on network connections, and no bandwidth restrictions. There are advantages to users too: data presented in one physical package is attractive, and users can be offered access to much higher-quality data than would be feasible over the Web.

There are a number of disadvantages of portable media for producers, many of them involving loss of control and information. With portable media, producers hand over their project data to the user in one convenient package; once released, the materials cannot be updated or augmented except by an additional publication. Furthermore, access control is difficult compared to networked resources, and producers cannot gather usage data or other information. Although they eliminate dependence on networks and hence can be valuable in environments where no network is available, portable media are vulnerable to loss, theft, damage, and physical deterioration. Their storage capacity—though increasing steadily—is nonetheless small compared to the size of a high-quality image archive, and is more suitable for delivering a focused sampling of materials than a large

collection in its entirety. From the development standpoint, too, there are disadvantages. Largely because it is often heavily customized, application development for portable media has traditionally been more expensive than for Internet-based delivery, and multi-platform development can be much more difficult (although there are also challenges in supporting multiple browsers and access tools for Internet-based distribution as well). Publication of portable media also poses potential difficulties; for one thing, it requires the production, packaging, storage, and shipping of physical objects, which may be part of a museum's ordinary work but requires a kind of staffing and space which libraries or academic projects may lack. And as few museums have access to media distribution and sales channels, their products do not get the shelf-exposure that CD-ROMs and DVDs released by both traditional and new media publishers achieve. As a result, although a small number of cultural CDs have sold over 100,000 copies, most sell only a few hundred and are only distributed in the local museum store.

Pros and cons of networked delivery

The advantages of networked delivery for producers center on ease of development and production: if the project is well planned, a single source of data avoids the need for duplication of data (and this is a consideration to be attended to carefully); there is transparent updating and fixing of application bugs; many client and server components already exist ready-made; and multi-platform development is far easier than for removable media. It is easy to create tiered access tailored to different target audience needs, and measuring usage is relatively simple. In addition, browsers provide a useful set of generic client capabilities. Institutions that subscribe to many digital resources—for instance, academic libraries—overwhelmingly prefer networked access over portable media because it simplifies their task immensely, eliminating the accession process, relieving them of the care of fragile physical objects, and broadening their users' access to the materials.

The disadvantages of networked delivery focus on the relative unreliability of networks and network applications: there are many potential points of failure; there are varying, non-standard, and buggy application implementations (Java, JavaScript, CSS, XML) and limitations in browser clients and interaction. Being on the Internet also raises security concerns, although these are being addressed in a number of ways. Charging for networked access is not as straightforward as the simple purchase of a CD-ROM, although again there are a number of workable models currently in use. Finally, the quality of the digital material that heritage institutions can make accessible via the Internet is constrained by the available bandwidth, and although bandwidth has been increasing steadily for the past two decades, so too has the demand.

For most purposes, some form of networked-based web application is a good solution, and this is the approach that most digital resource providers are now taking. Given the nature of digital collections, being able to update and add to the resource dynamically is a huge advantage for most content providers, as is the ability to add users effortlessly without producing and mailing additional disks and packaging. The Internet provides a direct distribution channel between the heritage institution and the user community. Most heritage institutions that use the Internet to make their digital assets accessible do so for free, perhaps because of current attitudes towards charging for networked information resources, although it is notable that licensed digital resources such as journals, full-text collections, and bibliographies are widespread and take an increasingly large share of library purchasing budgets.

Modes of access

Assuming a web-based, networked delivery scenario, the next set of choices concerns the application model for web delivery, and the underlying technologies and data structures that support the digital publication. As with other decisions, these will be strongly influenced by the scale and purpose of the project or program. The needs of a small-scale project with limited, unchanging content will be very different from those of a large organization with high-volume delivery needs.

The simplest form of web delivery is a static HTML page: essentially a document that has been encoded so as to be readable using a web browser. Pages of this sort reside on a web server and are delivered to the user for viewing when requested. They may be searched in very basic ways by various means: by using a search engine installed locally, or by using one of the available commercial search engines such as Excite or Google. Such pages may be quite simple to create and maintain, using standard HTML authoring tools such as DreamWeaver or Microsoft Front Page. Since HTML is a very basic markup language, it is readily understandable and easy to learn and use. Furthermore, the software required to publish such pages is simple and easy to support.

Some of the limitations of this approach have already been indicated in Section V, Digitization and Encoding of

Text; HTML has extremely limited descriptive capabilities and can represent very little of the intellectual structure or content of a document. Thus, while HTML is entirely adequate for the delivery of a project web site (home page, project description, contact information, background, and the like), it is not appropriate as a way of capturing the actual intellectual content of a site (primary source documents, metadata, bibliographic records, image collections, and so forth). But in addition, the structure of such a site—a simple collection of HTML files—does not offer any means of managing the data systematically, querying or displaying it powerfully, or maintaining it as it grows.

These two limitations—lack of descriptive power in the encoding, and lack of overall data management—can be addressed separately or together, depending on the project's needs and the nature of the data. The former can be addressed by adopting a more powerful descriptive markup, using an SGML or XML encoding system such as TEI, EAD, or one of the other high-quality encoding systems now available (see Section V for more details). There are now a number of free open-source tools, as well as commercial systems, for publishing XML documents on the web, using XSLT (Extensible Stylesheet Language Transformations). If the content to be published consists of textual items with rich internal structure that may be of research interest to users, XML is an effective way of representing this information and exploiting it in the user interface. Good examples of this kind of content are scholarly editions or full-text collections of historical or literary documents. Similarly, if the content consists of highly structured data such as bibliographic records, finding aids, or metadata, XML may also be a useful way of capturing and publishing it, albeit with a different kind of DTD. In all of these cases, you can use XML tools to create a publication which enables users to query the information in detail, view and manipulate the results, browse documents in their entirety, and perform other central research activities. At the moment, the requisite publication tools require considerably more technical competence than those for basic HTML; the XML markup is likely to be more complex, and the design and implementation of the XML delivery system require an XML programmer. However, there is a strong demand for XML publication tools, that allow a comparatively non-technical person to manage and publish XML documents effectively, and such tools may well start to appear within the next few years.

The second limitation, the need for more powerful data management systems, needs to be addressed at a deeper level: not in the encoding of the individual items, but in the creation of a powerful infrastructure in which they are stored and managed. Typically such systems are built on a database model, treating the individual items as database records and exploiting their regularity and structural similarity to allow for effective searching and manipulation. Such systems may be quite small—for instance, many web sites are based on small-scale database applications such as Access or even FileMaker, and for very limited purposes (such as publishing a small bibliography) these might well be an improvement over static HTML pages. However, for large institutional projects, robust database systems such as Oracle, Informix, and MySQL (the latter is free and open-source) are more typical and are capable of managing very large-scale collections. What the database architecture provides is the ability to query large numbers of documents with maximum speed, and to perform the kinds of processing —sorting, statistical manipulation, systematic comparison—that database tools are best at. Such a solution is ideal for a large collection of similarly structured items, where retrieval speed is very important: for instance, large metadata collections, image libraries, historical census data, and the like.

Using a database architecture offers a number of advantages in addition to power and scale. Once the underlying design is complete, the database can grow almost without limit as new records are added, without any work other than that of entering the new data. The system can include an input form which allows non-technical staff to make updates or additions, and the form can be made accessible over the web to accommodate off-site staff or collaborators. The newly added information may become live immediately, or can be marked as provisional, to be published pending review by an editor. Such systems also offer powerful workflow management and tracking possibilities which can be a valuable investment for a long-term project or one involving several participants.

One of the most distinctive features of both the XML and database-driven approaches is that in both cases, the web page that the user actually sees is generated dynamically from the source data (the XML document or database records), rather than being stored statically on the server. In the case of the database, the publication system retrieves information based on a user query, and then uses a style sheet to order, format, and display the results in HTML so that it can be viewed with a web browser. Similarly, although XML documents themselves are stored in a stable form, the XSLT stylesheets which turn them into HTML for web viewing can also perform more powerful transformations, such as reordering data elements, adding text and graphics, selecting and displaying small segments of the document, or even converting the information to another data format entirely. These dynamic capabilities open up a huge variety of interface options which go far beyond what is possible using simple HTML.

Finally, as suggested above, these two approaches—XML encoding and database architecture—can also be

combined to leverage the power of both, and indeed there are signs that the two technologies may be converging, as database systems become more complex and XML tools become more powerful.

In addition to the questions of technical implementation and user access discussed above, access models are also concerned with issues such as security, limitations on access and charging, which will be considered next.

Most of the projects interviewed in the survey for this Guide are currently providing free access to their collections, but many are looking at the alternatives or have already chosen a licensing model. Models for charging for the use of digital assets over the Internet are still not as widespread and straightforward as they might be (evidence of this comes from the continuing battles over Napster-type applications). But considerations such as protection of rights, or the need to generate revenue from digital assets, provide increasing motivation for projects to limit access to their online resources. These limitations can take several forms which may be more or less burdensome to the user. Typically users need to identify themselves before gaining access, via an authentication system of some kind. Such systems may involve direct user action, such as entering usernames and passwords, but there also exist modes of authentication that are invisible to the user, such as IP address authentication or access via an authenticating gateway. Whatever method is chosen, restricting access can be costly, determined hackers can find loopholes in most systems (particularly if the restricted data is valuable enough), and authentication systems may require a high level of maintenance and technical support. Projects or programs need to be sure that the revenue generated will justify these costs and that proper account is taken in the project planning of the responsibilities and legal liability of the agency with control over the server that distributes the assets. The advantages of free access to digital materials are therefore not just altruistic — there could be a significant overhead associated with limiting access. Many of the projects and programs interviewed were looking at revenue generation as a means of sustainability, so it is likely that secure systems that are easier to implement will be generally available in the near future. Corbis and Getty Images both believe that future access to picture libraries will be via the Internet, and they are building libraries and systems to make this happen. Many cultural institutions have generated income from their existing collections through the licensing of analog copies, and they need to make the shift to digital delivery of the material.

Future Trend:

XML

The advantages and disadvantages of static and dynamic web pages have been outlined above. One possible distribution method that combines the simplicity of authoring and hosting of HTML with the scalability and structure of dynamic pages is XML as a data format in a native XML database or content management system.

For further information see Ronald Bourret's sites:

http://www.rpbourret.com/xml/XMLDatabaseProds.htm

http://www.rpbourret.com/xml/XMLAndDatabases.htm

Structural Metadata

As the digital representation of our cultural and heritage material increases and becomes more complex the relationship of individual digital objects to each other, the item from which they were derived, the collection to which they belong and the way in which the digital objects are stored and organized will become increasingly important. It is this information that will enable future users to retrieve not just a faithful representation of an object but reconstruct, navigate and understand the whole context in which the object was created and used. For this to be achieved distribution systems will have to hold increasing amounts of structural metadata. This may also suggest holding such metadata in object orientated rather than flat file or relational database systems.

Metadata Harvesting

As digital objects, catalogues and finding aids proliferate on the World Wide Web, effectively searching and retrieving information across multiple servers, systems and domains becomes both increasingly important and increasingly difficult. This issue is a central concern for cultural heritage and humanities institutions, which share the goal of broadening access through digitization.

One solution to this challenge is Metadata Harvesting. Simply put, this is a protocol that exposes metadata on the World Wide Web to enable searching across repositories. The best known system is the Open Archives Initiative Metadata Harvesting Protocol (MHP).

Further details:

A non-technical introduction to MHP:

Clifford Lynch, "Metadata Harvesting and the Open Archives Initiative," ARL Bimonthly Report 217 (August 2001): http://www.arl.org/newsltr/217/mhp.html

and

Donald Waters, "The Metadata Harvesting Initiative of the Mellon Foundation," ARL Bimonthly Report 217 (August 2001): http://www.arl.org/newsltr/217/waters.htm

The OAI MHP protocol: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html

MHP tutorial: http://library.cern.ch/HEPLW/4/papers/3/

CIMI Working Group: http://www.cimi.org/wg/metadata/

CLIR Metadata harvesting project: http://www.clir.org/activities/details/metadata-docs.html

DLF and Metadata Harvesting: http://www.diglib.org/architectures/mdharvest.htm

University of Illinois at Urbana-Champaign Metadata Harvesting services: http://oai.grainger.uiuc.edu/

Conclusion

In determining the best delivery options, user needs should be balanced against producer advantages. Though portable media are cheap, reliable, relatively easy to sell, with no bandwidth problems and appear easy to archive, network delivery is generally easier and cheaper to produce and update. Tiered access for different audiences and for measuring usage is simple online. Despite bandwidth restrictions and security concerns, for most purposes networked delivery is the most common solution, providing a direct distribution channel that often reaches unexpectedly wide audiences.

Depending on the scale and purpose of a project or program, a database-driven system will generally be preferable to static HTML pages. Although new developments, such as metadata harvesting, are making content stored in databases more easily discovered, HTML pages are still more easily searchable from outside than dynamically generated pages, although within the resource itself the reverse is true. XML pages, though more costly to develop, are more economical and easier to maintain and update, and they offer vastly improved possibilities for data description, manipulation, and reuse. As with other decisions, this one will also depend on how users use your material as well as the scope of your project.

Access models also have security and economic aspects. Rights management and the need to generate income from digital assets may mandate access limits via identification and authentication systems, which, however, add another layer of cost for development and maintenance. New security and economic models are important to

watch for.

---

[1] Because of the way data are encoded when written to tape and the way error correction works we do not recommend that information be written to tape only in uncompressed format.

[2] Zip and JAZ are proprietary systems manufactured by Iomega.

[3] Note the 250MB drives cannot read 100MB media and vice versa.

[4] Hypertext Transfer Protocol: A set of rules that enable documents to be used on the worldwide web and viewed by browsers such as Internet Explorer or Netscape.

[5] File Transfer Protocol: enables files to be transferred from computer to computer using a simple program such as WSFTP or Fetch.

[6] Internet service providers most typically provide internet access for users. But many of the larger ISPs also offer server services which can accommodate the publication of large web-based resources, for a fee. If your institution is not able to maintain its own web server, or if the resource you are creating is too large to be published from your local servers, an ISP may be an option worth considering.

table of contents previous chapter next chapter

valid xhtml 1.1

XI. Sustainability: Models for Long-Term Funding

Introduction

Regardless of the quality and robustness of the digital resources created, they will not last long, let alone grow and survive in the long term, if the projects that produce them cannot be sustained. Sustainability thus refers to all of the considerations that go into maintaining the institutional context for the creation and maintenance of digital objects and resources, and supporting their long-term viability. Preservation—the maintenance of the digital objects themselves—is a distinctly different but related topic, dealt with in Section XIV. Digitization of cultural objects is still relatively new; the majority of projects and programs interviewed are less than ten years old. We are not yet able to take a long-term view of the life cycle of a digitization project, hardly even a medium-term view. In most cases, energies are naturally channeled towards start-up and building a critical mass. Although there was some recognition among projects and programs that a strategy for sustainability was desirable, it is probably true to say that sustainability in the long term is still generally pushed down the list of priorities by more immediate and pressing concerns. Readers should also consult Section II on project planning, which puts these issues into the context of the larger planning process.

Few projects have identified a point when, with their work complete, they will close down and pass the content they have created in digital form over to others for long-term maintenance and delivery. And yet for many projects, whether they plan for it or not, this is the likely outcome, either because of loss of funding or simply because the project has finished its work and has no further reason to continue. It is worth considering at the outset whether the digitization work you are undertaking has a finite scope, or is likely to continue indefinitely. Projects that are defined by the scope of a particular collection or author should plan for an appropriate exit strategy when their work is complete, including plans for ongoing delivery of the digital resource created.

The Survivors of the Shoah Visual History Foundation, for instance, advises that project managers plan out the whole system before any digitization takes place, envision the entire process and life-cycle of the project, and plan for it up to and including preservation of the materials. This plan should ideally include contingencies for re-locating the digital objects if, for example, funding declines to a point where it is no longer viable to maintain the materials.

Projects whose mandate is broader will be limited (or not) by other factors, such as institutional priorities and funding. For some institutions, such as national libraries, it is entirely appropriate that they should not have an exit strategy: the mission and funding of such institutions is consistent with long-term support and maintenance of the digital materials they create. But it is not only the national libraries that have such far-ranging plans: some large academic institutions are also making a commitment to keep digital materials over the very long term. The Robertson Media Center at the Clemons Library, and the Special Collections of the Alderman Memorial Library, both at the University of Virginia, intend to keep their digital deliverables viable and available indefinitely. The University of Virginia Library as a whole is launching a Central Digital Repository that will, in theory, become the sustainable home for projects currently delivered via one-off systems at the Robertson Media Center and Special Collections. The Harvard University Library Digital Initiative (LDI) adds another aspect to the strategies mentioned above. Repository policy strongly encourages deposit of at least one version of the digital object in an approved preservation format to ensure that new deliverables can be generated affordably in the future.

Most projects reported that loss of the digital materials would matter greatly—institutionally and to their user communities—but few could be confident of the continued stream of funding and other support that would be necessary to maintain their digital assets. There are high costs associated with ensuring long-term access, for example: regularly updating the interface, maintaining backup copies, migrating the data to new platforms and formats, and supporting the use of the resource, let alone providing for its continued incremental development over time.

Programs and projects need to think more strategically, and need to adopt a longer-term view by planning out the development of sustainability plans at the outset of their initiative. Funding agencies have a role to play in

insisting that provision for long-term sustainability be factored into grants They also have a wider responsibility in recognizing the cultural value of digital objects and contributing towards their sustainability. However, long-term sustainability is difficult to achieve on grant funding alone, and a reliance on “soft” money puts a project on a very insecure footing: not only because of the uncertainty of funding from year to year, but also because so much of the project’s resources must go towards the ongoing search for support. If funders value the digital cultural objects they can help maintain their longevity by encouraging the establishment of longer-term, strategic funding plans. Needless to say, the prevailing economic and political climate is relevant and not necessarily conducive to a strategic, long-term view.

In this section we will look first at the funding issues in more detail, and then discuss planning strategies and methods of sustaining resources through use.

Funding

Central Funding

The Library of Congress’s National Digital Library Program (NDLP) intends to keep its digital deliverables available for as long as possible. The adoption of an exit strategy, in case the Library found itself unable to meet the costs of sustaining the NDLP’s digital assets was considered unthinkable. On the contrary, the Library recognizes that it is more likely to provide an exit strategy for other institutions that find themselves unable to sustain the digital resources they have created.

For the Library of Congress, the sustainability of its digital resources would not depend upon their generating income streams to make them self-sufficient. In our interviews, the Library reported that it had already obtained some funding towards providing a level of sustainability for its digital assets and is working to make these resources visible to national audiences, thereby building a large group of invested stakeholders.

Similarly, the National Library of Norway’s overall statement of intent, or mission statement, refers to the need to establish a framework for digital libraries and collections that make information available to everybody, wherever they are, in a democratic way. The emphasis is on access and inclusion. They intend to establish a long-term digital library with core functionality, and then provide different “windows” onto this information. As with the Library of Congress, there is no exit strategy — on the contrary, the digital activities at the Library aim at building a sustainable digital library, with sustained and secure funding.

Museums face the same challenges in finding opportunities to fund digital collections project–by project and collection–by collection. While the Berkeley Art Museum/Pacific Film Archives builds its digital projects with sustainability and preservation in mind, it must fund the creation of new digital collections in exactly this way. Early digitization efforts at the Peabody Museum of Archaeology and Ethnology at Harvard University also followed this methodology.

Base vs. project funding

Many projects, like the Cornell Institute for Digital Collections (CIDC) at Cornell University have benefited from very generous lead-gift funding and continue to rely on soft money. However, the use of grants and gifts to fund digitization initiatives means that the prioritization and selection of material for digitization may be driven by an assessment of those areas most likely to receive funding at any given time (through their timeliness or appeal to a particular funder), rather than by an overall digitization strategy. These projects recognize the need for a stable funding environment.

Similarly, the Virginia Center for Digital History (VCDH) at the Alderman Memorial Library at the University of Virginia (UVA) has benefited from funding from a wide variety of sources including UVA, National Endowment for the Humanities (NEH) and private donors. The main problems with the funding have been uneven distribution and inconsistency. With the benefit of hindsight, the Center felt it would have been better placed if it had had solid base funding with project funding added on top.

Securing longer-term funding

There are three primary ways to secure long-term funding for digital assets: commitment by the originating institution to meeting the costs associated with the digital asset, endowment funding, and making the resource

self-sustaining. Of these, endowment funding is the most difficult to obtain and none of the projects interviewed for this *Guide* described a successful approach of this sort.

Several projects, however, have set in place a funding model that relies on generating ongoing support from the resource itself. The Brown University Women Writers project is dependent upon self-generating funds; it has looked for endowment funding but now hopes that the resource itself will generate sufficient income to meet its long-term costs. The sustainability of the Colorado Digitization Project also depends on self-generating funds and it is considering establishing a membership-based organization to generate revenue in a coherent and consistent way. In making its resource self-sustaining, JSTOR has established a basic fee structure for participating institutions: a one-time Archive Capital Fee (underwriting digitization costs and an archives reserve fund) and an Annual Access Fee (supporting ongoing access to the database and the addition of new volumes to the database). JSTOR director, Kevin Guthrie, comments that nonprofits should develop revenue sources that "match the nature of the mission-based uses of their funds." As JSTOR's work of archiving journals continues, so stable and recurring funding is essential. JSTOR is thus building a reserve fund, the proceeds from which will help fund the ongoing cost of archiving.[1]

Some plans for self-maintenance face greater challenges. The Thesaurus Musicarum Latinarum (TML) project is dependent for long-term sustainability on self-generating funds, but the TML has not yet secured these resources. Should resources prove insufficient, the exit strategy would be to break up the TML and the larger consortium group, and move them elsewhere. Another possibility would be to publish the TML from a private server. The TML experience suggests that projects need to recognize that even institutional support is tenuous, that digital data are easily destroyed, and that some projects will not survive.

Cooperation and joint initiatives with other types of institutions or agencies may also provide a route towards longer-term funding. Where several institutions combine to support the creation of a resource, the burden for each may be quite manageable. Similarly, funders may look more favorably on joint efforts that show a broad institutional commitment to the resource. The CIDC at Cornell, for example, has cooperated with libraries, museums, and archives, at all levels from local to international. While these are useful processes, cooperation is not without problems and firm rules have to be agreed to produce a successful project. However, cooperative agreements of this sort can provide a route to funds that are otherwise inaccessible. A more detailed discussion of inter-project collaboration and its benefits and risks can be found in Section IX.

Effective strategies and planning towards sustainability

The Digital Imaging and Media Technology Initiative (DIMTI) at the University of Illinois is an excellent example of "joined up thinking" in relation to digitization. The benefits of establishing a body such as DIMTI at the outset of digitization are demonstrated in the strategic overview that it enjoys. The fact that it is a single body with responsibility for dealing with generic issues from the outset has advantages in creating a unified institutional vision. Furthermore, DIMTI demonstrates that such a body does not need to be large when supported by other appropriate administrative or managerial bodies. The fact that the University of Illinois Library has a relatively flat structure seems to have been beneficial in this case.

Although various forms of base and soft money remain the predominant forms of sustainability, the projects surveyed for this *Guide* tended to be large and relatively well established. Therefore, they may be generally less concerned about their ability to secure future funding than smaller and more recent projects. However, some projects have developed alternative models for sustainability.

The licensing model developed by the JSTOR project secures long-term funding by creating durable relationships with other stake-holders: most significantly, the collection's current and future users. In this model, access to JSTOR collections for UK higher education institutions is licensed via the UK Higher Education Funding Council's Joint Information Systems Committee (JISC). The JISC succinctly summarizes the benefits to both parties as follows: "the agreement between JSTOR and the UK Funding Councils (acting through the JISC) is a twenty-five year Agreement commencing on 1st August 2001 and terminating on 31st July 2026. The length of this agreement is to reflect the on-going partnership between JSTOR and the JISC, to allow institutions to feel confident in long term access to JSTOR collections, and to underscore the value this resource brings to the UK academic community." Such sub-licensing agreements are effective methods for projects that have a large and identifiable institutional audience to secure funding and sustain the project through long-term use.

The Colorado Digitization Project's collaborative, not-for-profit model is a second alternative. The project has a broad range of funders (nine in total) and participation is open to libraries, library systems, archival organizations, selected pilot schools, historical societies, and museums. The involvement of more than 50 collaborative

projects and four regional scanning centers, and the emphasis on sharing and transferring expertise means that sustainability does not depend on a single funder or institution but is shared by the collaborative partners.

Neither of these strategies is guaranteed to ensure sustainability. Licenses may not be renewed or obtained in sufficient numbers and contracts for large consortia, particularly regarding ownership of material, are difficult to achieve. Nevertheless, the demand for grant-based funding will always outstrip its supply and such models indicate alternatives that can go some way to achieving sustainability.

Focusing resources and explicit statements of intent

Focusing resources on a clearly identified target audience has been shown to be helpful in sustaining projects, as that audience has an important stake in the future sustainability of the program. The UVA Electronic Text Center reported that the most effective dissemination strategy for building a sustainable resource was predicated on the Center's local focus on research, teaching and learning, and the grounding effect this produces. Similarly, the William Blake Archive (also at UVA) has produced an explicit statement of intent that covers its rationale, scope, significance, primary audience, long-term sustainability, level of faithfulness to the originals and suitability for different target audiences. This level of focus has enabled it to sustain a high degree of momentum and hence of attractiveness to its audience, which in turn strengthens its funding position.

Sustaining resources through use

Ensuring internal and external use of a resource can contribute to its sustainability. The Walker Art Center (WAC) and Minneapolis Institute of Art (MIA) both intend to ensure that the digital process becomes part of their institutional culture, ensuring that all types of user use the resources from day to day. They will educate staff to identify and use the objects in everyday work, and if they are successful, this integration of digital materials into the mainstream of users' processes will help ensure the sustainability of the digital objects. Similarly, DIMTI is an example of a program that has set itself the goal of integrating digitization into existing Library activities and structures; this is a far-sighted and effective mechanism to garner institutional support and maintain digitization activities in the long term. The sustainability of digital resources will be increasingly assured if academics are successfully encouraged to use them.

Summary Box:

Key elements in establishing sustainability:

- Effective strategies and planning towards sustainability — ensuring that information about your resources is easily available and that potential users are encouraged to adopt the resources (see Section II, Project Planning);
- Use of appropriate standards to facilitate the sustainability of resources by making them less susceptible to changes in technology, and easier to migrate (see Section XIV, Preservation);
- Developing groups that are stakeholders in your resources at the outset (see Section IX, Working with Others);
- Getting into the mainstream — by ensuring the widespread use of digital resources in learning and teaching as well as research (see Section II, Project Planning);
- Migration of skills as well as technology.

Migration of skills as well as technology

The Genealogical Society of Utah, Family History Department reported that it is not only technology that ensures the longevity of materials, but also migration of hardware, software, media and, importantly, people skills. The GSU is constantly reviewing its management procedures to ensure that no information can become obsolete due to hardware, software, media or skills being lost. (See the discussion of skills in Section II, Project Planning.) Nearly all projects will eventually need to migrate their digital resources to new environments and will require funding to do so, but these costs are rarely built into the project's base funding.

Building and sustaining digital collections: NINCH/CLIR models

Cultural heritage institutions face a number of challenges when placing digital collections online. Not least of these is the problem of how collections can be built and sustained in harmony with the sector's non-profit culture. In February 2001 the Council on Library and Information Resources (CLIR) and the National Initiative for a Networked Cultural Heritage (NINCH), supported by the Institute of Museum and Library Services, convened a meeting to discuss what sustainability models were available to the sector. The meeting included representatives from museums and libraries, legal experts, technologists, funders and web enterprises.

In discussing business models the group concluded that whether the digital collection is for profit or not, large amounts of capital, new skills and new organizational cultures are required. Furthermore, non-profit organizations must adopt the same rigor for accountability, efficiency and value for money as commercial enterprises. The latter may have a competitive advantage in their ability to identify, test and target audiences, but any successful project must have clearly identified its purpose, mission and audience. Nevertheless, non-profit organizations generally have a longer period of time to realize their goals, and the expectations of trustees and governors can differ considerably from that of shareholders or investors. For the report, *Building and Sustaining Digital Collections: Models for Libraries and Museums*, see http://www.clir.org/pubs/reports/pub100/pub100.pdf

It is becoming apparent that web access to digital collections is not just broadening access but providing access in new ways. This presents a challenging dichotomy for non-profit organizations. How can the convenient and unrestricted access to digital collections that users expect be reconciled with the need to recover costs? Moreover, the resources, outlook and skills required to build a digital collection are somewhat different from those required to distribute it. The non-profit organization may find that the latter are far harder to integrate into their larger mission.

Cultural heritage organizations could license their digital content to a commercial distributor, but it is by no means clear that the analog/digital divide is the point where non-profit organizations cease to owe users free access. Even though access to analog collections has never been entirely free, it is still uncertain at what point —and to what extent—cultural organizations should start using web delivery to load costs onto the user.

An alternative model is strategic collaboration between non-profit organizations. (One such model examined at the CLIR/NINCH meeting was between the George Eastman House, Rochester, New York, and the International Center of Photography in Manhattan. See page 7 of the report cited above). The costs are not necessarily going to be lessened through this approach but it does provide a means of spreading and managing the risk involved, building economies of scale and meeting the need for standards and practices for interoperability and access. However, as noted above, even between large organizations such collaboration is no guarantee of a successful return. Considering the infrastructure and skills required for creating and delivering digital content, small and medium sized libraries and museums are likely to find that collaboration with a larger partner is the only method available to them.

Despite the risks and costs involved, participants at the meeting voiced the opinion that not to enter the digital arena at all was itself a greater risk. At least the risk could be shared by working collaboratively. Furthermore, the burden of expectation might be diminished if funders viewed as experimental some ventures where the return might be non-existent or elusive.

The elements of sustainability identified at the meeting require further work by the sector. Participants emphasized the need for standards and best practices, coherent and common digital architectures, and a means of creating and sharing knowledge. In order to deliver these desiderata, the group suggested a number of ways forward, including guides to best practice, a clearinghouse for technical information, a registry of digital conversion projects, digital service bureaus, regional centers for digital content production, and “tool kits” for data and communication design and development.

---

[1] “Developing a Digital Preservation Strategy for JSTOR, an interview with Kevin Guthrie,” RLG DigiNews (August 15, 2000) http://www.rlg.org/preserv/diginews/diginews4-4.html#feature1. See also the section “Who pays what,” in Dale Flecker, “Preserving E-Journals,” D-Lib Magazine (September 2001) http://www.dlib.org/dlib/september01/flecker/09flecker.html

valid xhtml 1.1
abp~03/03

XII. Assessment of Projects by User Evaluation

Introduction

There are two main reasons for carrying out evaluations: to ensure that digitization initiatives produce materials that meet user needs, and to assess whether the original objectives of the project were achieved. Throughout this *Guide* we have stressed the substantial and long-term financial investment that digitization requires, and have drawn attention to the demands it places on limited organizational resources. It makes good sense, wherever possible, to assess the results methodically and make the benefits of this investment evident.

Evaluation provides a mechanism to measure objectively what has been achieved through digitization. Funding bodies increasingly require that digitization projects deliver an evaluation and impact assessment report as part of their final report. In addition to helping projects to be more accountable to the user community, evaluation provides an opportunity for projects to discover who is using the digital collections and how. We make many assumptions about our users and their needs, often basing our expectations of digital usage patterns on existing usage of analog resources. However, at best, these can only be educated guesses, since access and usage can change completely in the networked environment. For instance, use of postcard collections has always been limited, but when made available in digital form their use rises dramatically; ease of access is the key. Only by carrying out evaluation with our users can we find out how the digital resources we create are actually being used. Despite the spreading use of digitization in the cultural sector, we know surprisingly little about its effect on the various user groups we aim to serve and the way digital collections influence their teaching, research, life-long learning, personal development, or entertainment. Research and evaluation will help indicate how we can enhance the functionality and usability of the digital resources we are creating.

At a more practical level, giving the target user group the chance to evaluate and test the delivery system, and in particular the interface, at an early prototype stage should enable the development program to take advantage of feedback and suggestions for improvements. This approach makes most sense when projects and programs have built in the resources needed for iterative prototyping, so it is a good idea to ensure that you have staff and time to implement revisions if you get early feedback from potential users. Commercial and cultural institutions report that this approach allows them to be more responsive to a broader spectrum of users with a variety of technical experience and familiarity with the subject. User evaluation and research can provide answers to these (and other) questions:

- Who uses the digital collections? What are their demographic characteristics, interests, level of computer skills?
- How are the digital resources being used? What navigation paths are followed? What are the popular choices, the types of searches performed?
- How does the use of the digital collections relate to that of the analog ones? Are they complementary? Is the profile of the users for each category different? Has usage of the analog resources changed as a result of digitization?
- What is the effect of the digital resources on the users? Are they satisfied with the service and collection? Does the resource trigger interest in new topics, or challenge existing perceptions? What are the formal and informal learning outcomes?

The questions you ask will depend on the aims of the program and the particular focus of your interests. For example, if you are interested in educational uses of the digital surrogates of a museum's collections, the main issues to explore might be the kinds of understanding and learning the digital resources can enable, and how you can ensure that visitors who use them also engage with the original objects on display. If the digital resources are to be made accessible in a museum gallery, then evaluation activity might examine how computer interactives affect the visitors' experience, and how they affect the social context of group visits and interaction between participants in the group. In an academic library setting, evaluation might assess whether digitizing a given collection is likely to address local curricular needs, and what kinds of improved pedagogical outcomes might result. Evaluation can also help discover how effectively digital surrogates represent the original. Do

visitors/users prefer to see the actual object or do the digital surrogates have the same impact? Questioning the user base can reveal this and any evaluation activity should approach this issue.

In making these assessments, remember also that many resources will be used by very different user groups: users from a variety of age groups, educational levels, backgrounds, and interests. Their needs and expectations will also differ considerably. An essential starting point before beginning any evaluation is thus to categorize the user community to make the assessment more informative. For example it might be useful to group users by:

- their subject area of interest (e.g., Classical Greek archaeology, Abstract Expressionist paintings, the history of Ethiopian Jews);
- their age group and educational level (K-12, college student, faculty, life-long learner);
- their function or role (e.g., librarians, conservators, curators, students, general public);
- the use they would make of the resource (e.g., research, teaching, life-long learning, casual browsing).

You can further refine this categorization by considering the potential user communities with reference to the project's defined context and priorities. For instance, it may be important to ascertain whether specialists actually use all the extra search facilities they say they want, or whether they use digitized resources more or less as the general public do. You may discover that different communities actually use the resources in very similar ways. On the other hand, certain users might have particular requirements, for example in relation to image quality or searching facilities.

Types of evaluation

Evaluation is classified into three types according to the point during the lifecycle of the project at which it is carried out: front-end, formative, and summative.

Front-end analysis is carried out before a program or application is developed. This type of evaluation can gauge potential users' reactions to the subject matter; assist in the selection of content areas and themes; provide feedback about the type of service and functionality that would be suitable; and enable the project to make a general assessment of the attitudes that stakeholders and users have towards the proposed development. Involving both groups early on helps ensure that the final product or service conforms more closely to their expectations and needs. *User needs assessment* often plays an important part in front-end analysis. This technique investigates the requirements of target users—(search tools, image or sound quality, depth of metadata—using methods such as focus group discussions or interviews, described below). The front-end evaluation process should also examine whether digitization is the best way to achieve the aims of the project or whether simpler and more affordable solutions might be more appropriate.

Formative evaluation takes place during the development phase and its results help refine and improve the design and delivery of the resource. It can be used to test the appropriateness and intuitiveness of the user interface and pinpoint problematic areas and programming errors. This is a vital step in the design and development of all digital resources. Even if the final product is not perfect, it will be better than if no user testing were carried out at all. It is never too early to test and involve future users in the design process. Even mockup screen views drawn on paper, sketchy web pages, or crude prototypes can provide valuable feedback and suggest changes before too much time and effort have been expended. Where a large user sample is too difficult to administer, even a brief survey with a small sample number of users will, in most cases, offer useful information at this stage. Formative evaluation also provides an opportunity to assess users' perceptions of the project's content. Different user groups might provide crucial feedback about the kinds and quantity of contextual information, metadata and tools for using the collection that they would find useful. The project will quite possibly discover at this stage that different groups of users have different expectations and may find that there is a need to narrow the focus of the project rather than attempt to provide the definitive resource to all potential users. Since in many cases digitization is a continuing effort for cultural institutions, formative evaluation should also be an ongoing activity, integrated into the digitization chain and implemented as part of the working processes.

Summative evaluation measures the effect and impact of the completed program or a distinctive stage of its development. It is good practice to define particular stages at which to conduct summative evaluation—for example, after a particular phase of activity or once a particular deliverable has been completed—especially where digitization is an ongoing activity. Summative evaluation is often more thorough and yields more determinate results, involving real users as opposed to targeted potential ones. Since it takes place after all the selected materials have been digitized and associated interface and tools designed, it offers a more accurate picture of how these are perceived and used than that provided by formative evaluation. When the whole range

of resources is made available, interesting observations can be made, for example, about the relationships between them, the most and least popular materials, and the associations users make between different materials. It may also be possible to find out why users find some materials of greater interest than others.

Often the summative evaluation is the first time that evaluators can measure in depth the effectiveness of interpretative exhibits and gallery kiosks, in relation to the surrounding space, and in the context of the exhibition itself. This is also the opportunity to explore the dynamics between real and surrogate objects, visitors, and computer interactives. The approach and tools chosen for summative evaluation will naturally depend on the aims of the survey and the reasons for carrying it out.

Evaluation methods and tools

Measuring and recording the effect of digital resources in cultural and educational settings, as well as the personal meanings that people derive from them, is a complex and difficult undertaking. Gay and Rieger (1999) have argued convincingly that "[m]edia- and technology-rich environments, such as cultural Web sites, demand equally rich data collection and analysis tools that are capable of examining human-computer interactions." There is no single golden method for evaluating digital programs and measuring their effectiveness and impact. Experience has shown that it is better to combine several methods (known as 'triangulation') in order to verify and combine data, relating quantitative with qualitative results. The dichotomy between the quantitative and qualitative approaches is usually artificial, since they work best in complementary ways, illuminating different aspects of a complex phenomenon.

Several useful pointers have emerged from experimental work in evaluation:

- Feedback from real users of digital collections and programs in their 'natural environment' (ethnographic research or the naturalistic approach) is crucial and usually more useful than testing in controlled situations with users solving tasks which evaluators have pre-defined.
- It is important to examine the use of the technology in its social context, rather than relying only on technological measures.
- As with user surveys in general, it is essential to define and follow a rigorous sampling methodology in order to acquire valid results that allow inferences to be made about the whole population of users. (Random sampling ensures that every person in the user population has an equal chance of being selected to complete a questionnaire or offer information, but this may not be the best method. For example, if you wished to gauge the effectiveness of the materials used by elementary school children, a stratified random sample might be more appropriate.) When collecting your sample, try to spread the process out over different days of the week and times of the day.
- Don't forget the non-users! Why do potential users not use your digital collections? Is it because they lack access to adequate technology or because the deliverable and the way it was presented did not appeal to them? What are the biases and barriers against the use of your digital collections? The answers of non-users can give you useful food for thought.

Some of the most commonly used methods in traditional evaluation work can be applied to digital programs, and these are briefly described below.

Computer logging of user interaction

Automated logging of user interaction with a digital resource provides a reliable way of recording users' choices and the path they selected through the website or program. There are numerous programs for recording web usage statistics, several of which are freeware or shareware.[1] Most of these offer possibilities for graphic displays, diagrams and categorization of results according to various parameters (e.g., requests by day, month, year). They usually record the number of requests made, the folders and files requested, and a list of countries or types of sectors from which the users come, often based on an analysis of the IP address of their computer. By measuring the length of time between selected links, researchers can estimate approximately the amount of time spent on individual pages. Although web statistics are notoriously unreliable and difficult to interpret (for example, estimating the number of individual users from the number of requests), they can still be very useful in a comparative way, even if the absolute figures should be treated with caution. In general, web statistics programs offer insufficient information to build a profile of the users beyond an estimation of where their computer is based. More detailed and customized information can be recorded, using specialized tools developed in JavaScript, Java, Visual Basic, or C. Research into scripting and computer logging is being carried

out at Virginia Tech (http://www.cs.vt.edu/research).

Once the scripting has been set up, computer interaction logging is generally an easy and objective way of obtaining a large set of data, which can be analyzed statistically. One problem is that when this method is used for programs in public access areas, it is sometimes difficult to differentiate between the interactions of different users. Another is that although it reveals the most popular user choices, it does not explain why they were chosen. The results are not very meaningful on their own, but can be useful when combined with interviews, focus group discussions and observation.

Sites that require user registration offer a richer range of possibilities for evaluation and contact with the users, although here evaluators need to take into account such issues as privacy and data protection (see Evaluation and Privacy Issues below). Registration requires users to register and then log in to the site using the same user name and password on each occasion. This provides more accurate information and can help develop trend data: for instance, to what degree use by a particular category of patrons may be rising or declining over time, whether particular types of material tend to be accessed by specific user groups, or whether certain user groups use particular types of material or explore the resource in particular ways. Although registration may deter some users, it helps if the registration page is carefully designed, with clear information about the purpose of registration and assurances as to how the data collected will be handled.

Techniques similar to web usage logging — and often more accurate — can also be used with stand-alone multimedia programs or CD-ROMs.

Electronic questionnaires

Although the results generally pose problems for valid statistical analysis, as the sample is self-selected, electronic questionnaires provide an easy way to obtain feedback from end users. They work by encouraging users to answer questions about the resource and their use of it, often by typing or clicking on multiple choice answers. Common practice is to require the user to make an active choice to complete the questionnaire, but more recently institutions, eager to better understand their users, have implemented short evaluation questionnaires that appear automatically a few minutes after the user has entered the site. Although more intrusive, this usually generates a high number of responses. The results of the questionnaires can be automatically loaded into server-side databases for subsequent analysis.

Questionnaires can also be sent by email or mailed to users whose email and postal addresses are known or who have provided this information. This allows for more flexibility and customization compared with the standard electronic questionnaire approach. Attractive and clearly laid out printed questionnaires placed next to the computer terminals can encourage local users to leave their impressions and comments about the program. Providing enough pens and visible boxes or assigned locations for returning the questionnaire can help increase the number of responses. Again, this method is not statistically valid, as it records answers from a self-selected and not necessarily representative sample. It might be a useful option for recording information about local users, as long as it is not the only method of evaluation.

Observation and tracking

Observing how people use digital collections in the classroom, gallery exhibition, reading room or public space can be very illuminating. It provides an opportunity to collect information about the physical, social, and intellectual contexts that affect the use of the digital resources, indicating, for example, the relationship with real objects in the gallery, the interaction between groups of users, or the ways students make connections between primary sources and the school's curriculum. Observations can be recorded on data collection sheets, with circulation paths or with checklists of specific behavior categories, together with personal notes. Video recording and web cameras offer alternatives and can produce a wealth of data, although these often take longer to analyze, thus raising the cost. Observation raises privacy issues, which we examine below.

Interviewing and focus group discussions

Interviews and discussions with a small number of targeted users provide an effective method of evaluation and offer the opportunity for both structured and open-ended data collection. Discussions with focus groups are often held during front-end analysis, as well as at other stages of evaluation. These are participatory sessions with

small groups of people, from the real audience or a particular targeted subset, who are encouraged to express their opinions about the digitized resources and how they are using or would like to use them. Interviews and focus group discussions can be open-ended, with the interviewer or discussion moderator interacting freely with users, or follow a pre-defined set of questions. Open-ended interviewing can be particularly useful for front-end analysis, to test how and what the targeted audience thinks about a topic before beginning program development. If an application is intended for specific groups (e.g., a researcher's resource or a schoolchildren's outreach program), discussions with focus groups can be very useful during the planning and development stages. These preliminary discussions often help outline a list of questions for more formal interviewing. Interviews usually provide many useful and meaningful data but are time-consuming to administer, demanding on the interviewer, and difficult to analyze and categorize. Projects should consider using ethnographic software for annotating transcripts of interviews with users. This software permits the researcher to aggregate comments with a given annotation label.

When testing a prototype, interviewing and observation can take two forms (often described as 'cued' and 'uncued' testing). Cued testing involves explaining to users what the program is about and asking them to perform specific tasks or to answer questions. Another possibility is engaging users in conversation and encouraging them to 'think aloud' as they go through the program, while recording their responses. With uncued testing, users are observed unobtrusively as they use the program and are then asked questions about their experience.

Checklist Box:

Checklist of criteria

Identifying appropriate criteria is vital for every evaluation. These depend on the scope and purpose of the study, the aims of the digitization program, and the time and funds available. This checklist outlines some of the basic aspects of digital resources that can be assessed. Each project must develop criteria that provide appropriate measures reflecting the goals of their own program and its users.

User interface/delivery

- Is the user interface consistent and appropriate to present the subject matter to the users?
- If icons are used for navigation buttons or commands, do the users understand them?
- Is the quality of graphics, images, sound, video adequate?
- Is the text legible (fonts, sizes, layout, spacing)?
- Are the media successfully integrated?
- Is the level of interactivity appropriate for the intended audience and environment?
- Are all the interface elements presented appropriately on different computer platforms?
- Can the program be used effectively by disabled users? Does it conform to ADA requirements for disabilities?
- Does the delivery of the program cater to different learning styles and types of audience?

Structure/navigation

- Is the structure of the various components appropriate to the content? (linear, hierarchical, network, combination)
- Is the resource easy to navigate? Does it indicate the user's position, prior moves, and available paths?
- Does the type or depth of indexing match the content adequately?
- Is the method of labeling content adequate?

Programming

- Are there any programming problems or errors (e.g. dead-ends and broken links)?
- What happens if the users do not use the application in the way it was intended?
- Does the application allow for user mistakes?
- Is there feedback given during processing operations that may take a long time?

Content

- Is the number of digital objects adequate?

Are the amount and depth of accompanying information about the objects adequate?
- Is the depth of indexing well matched to the content?
- Is the information accurate?
- Is the information complete or are there important parts missing?
- Is the information about the whole project, program, service or the individual collections appropriate, correct, and clear?

Overall impressions

- Does the resource provide useful information, arouse curiosity and interest?
- Is it easy to use?
- Does it attract users? What type and how many are using it?
- Does it hold attention? How long are they using it for?
- Does it fulfill its intended purpose? (e.g. does a research resource assist researchers effectively? Does an interpretative museum application help users get to know and better understand the collections?)

Link Box:

There are a number of resources providing guidance in the planning and design of evaluation strategies. Some of these are listed here:

- The Institute of Museum and Library Services (IMLS) in Washington, DC has a site on 'Outcome Based Evaluation', which offers an overview of this type of evaluation with other resources and links: http://www.imls.gov/grants/current/crnt_obe.htm.
- Professor Bill Trochim's Center for Social Research Methods (http://trochim.human.cornell.edu/) at Cornell University and the American Evaluation Association (http://www.eval.org/) have websites that offer useful information on evaluation and social science research methods.
- Jakob Nielsen's site on Heuristic Evaluation (http://www.useit.com/papers/heuristic/) offers information about the evaluation of a user interface design.

Evaluation and privacy issues

Projects should be aware of the IRB (Institutional Review Board) protocol, designed to safeguard rights and welfare of human research subjects. The main points to consider are:

- Risk to subjects
- Selection of subjects
- Informed consent
- Safety of subjects
- Privacy and confidentiality

Full guides can be found at http://ohsr.od.nih.gov/info/einfo_5.php3.

Some of the evaluation methods proposed may be seen as an invasion of privacy. In the case of observation, however, if you inform all the users and seek their permission in advance, experiments show that their behavior is affected and the results skewed. The evaluation team will have to address this issue and decide on the most appropriate way of notifying users or addressing the problem in general. Having policies in place to ensure that data collected are used sensitively and anonymously can also help to address some of the ethical concerns. In addition, evaluators need to be aware of the privacy protections afforded by federal and state laws.

Evaluation can help us realize the full potential of the technology to create powerful and attractive applications that assist users to understand digital resources in meaningful and relevant ways.

Who should conduct evaluations?

A final question is who should carry out the evaluation work itself. Evaluation work requires time, resources, and

skilled staff. It can either be conducted in-house or contracted to external professional survey or marketing companies, as is often the case with quantitative work, or to evaluation experts. Each option has different implications, advantages and disadvantages.

Carrying out the evaluation in-house can help to reduce costs, ensure that the work is carried out by staff members who are familiar with the mission of the organization, control the design of the evaluation and the way it is conducted, and provide for continuity between the evaluation and the way the organization takes advantage of the results. Before beginning, it is essential to ensure that staff have adequate training and expertise in designing and carrying out evaluation work and analyzing the data that are collected. For example, quantitative evaluation might require training in sample design and statistical analysis, while training in moderating focus group discussions might be useful for qualitative work, and system design skills may be essential for developing computer-based evaluation systems. Libraries, archives and museums may collaborate with specialists in information studies departments or with assessment professionals at area universities or colleges who can provide them with access to the necessary skills. One difficulty is that evaluation work is very time-consuming, and using in-house staff to carry it out will inevitably divert them from other activities. If you lack the skills in-house then it will prove more cost-effective to outsource the activity. For large-scale studies, that may entail drafting an RFP (Request for Proposals). The RFP should scope the problem that you hope to address through evaluation and ask the respondents to outline the methods they would use to conduct the evaluation work and how they would propose presenting the results.

Conclusion

Evaluation, like project documentation, is often viewed as ancillary to the main project goals, a separate task which takes a back seat to the work of creating the digital resource itself and disseminating it to users. However, as this section has suggested, evaluation not only offers tangible benefits in improved design and better responsiveness to user needs, but also may help avoid disastrous missteps in the planning stages, while change is still possible. Evaluation also enables you to document your project's effectiveness and make a persuasive case for new or continued funding. Careful planning will ensure that evaluation does not get in the way of efficient project development, or consume a disproportionate share of project resources. Conducted in this way, evaluation is an extremely valuable component of digitization work.

---

[1] For example, see http://www.ics.uci.edu/pub/websoft/wwwstat/, http://www.extreme-dm.com/tracking/, or http://www.analog.cx/

valid xhtml 1.1
abp~03/03

XIII. Digital Asset Management

Introduction

The *Guide* has already detailed the process of creating and distributing digital collections. But the process of digitizing material does not merely create an intellectual object: it creates a valuable asset, and consumes considerable resources in doing so. This section will look at mechanisms by which the institution that created or holds these digital assets can manage them to maximum advantage. It explores the value of the digital products to the organization, how they should be treated and managed in order to maximize their use, and how the institution can ensure that it receives full benefits from the sizeable effort and costs consumed in their creation. It describes the digital management options and assesses the need for digital asset management structure.

The digitization process produces digital assets whose significance is equal to any other asset that an organization may hold, such as computers, manpower or intellectual content. Indeed, they have been described as 'working' or intellectual capital. Just as an organization seeks to make efficient and effective use of its financial, manpower, and natural resources, it will now wish to use its digital assets to their full potential without reducing their value. In the digitization initiatives that were surveyed as part of our research, the digital creations will be presented and offered to the end-users in the same way as any other commodity is made available to customers. However, when the digital product is taken or 'bought', the organization still retains the product itself and does not need to engage in further manufacturing to reproduce the product. If the organization manages the product, for example through licensing (see Section IV on Rights), once the customer has purchased the digital material, it effectively remains an asset that the holding organization can exploit in perpetuity, provided that the value of the digital asset is not compromised.

The need for digital asset management

As the size of digital collections grows, so do the financial resources that are consumed in their creation. The costs of managing the resources also increase, although not necessarily in direct relation to the growth of the collection. It therefore becomes imperative to manage the resources effectively, a process that is commonly referred to as Digital Asset Management (DAM). There are many reasons for initiating the use of DAM.

Digital resources are often viewed as ephemeral and fragile, while at the same time they are seen as objects that can be easily recreated — one only need scan the picture again. Fortunately this lackadaisical attitude appears to be diminishing in most organizations and in its place is the growing belief that digital resources are, at the very least, as valuable as the time, effort, and finance that has gone into their creation. At the Berkeley Art Museum, Günther Waibel considers the move from taking transparent film surrogates of the Museum's collections to taking surrogates with a digital camera to be a critical one. The digital objects become 'mission critical institutional assets' that require management strategies similar to those already in place for the Museum's existing collections (Waibel 2000). Digital Asset Management allows organizations to maximize the use of these resources, ensuring that their value is maintained, while generating institutional savings.

Digital Asset Management (DAM) involves:

- Creating an efficient archive that can hold digital resources (such as images, audio and text) and the metadata that describe them;
- Implementing an infrastructure to ensure that these electronic data are managed and preserved in such a fashion that they will not become obsolete;
- Implementing search facilities that enable users to identify, locate and retrieve a digital object.

The benefits of implementing DAM include:

- Centralizing discovery and access;

- Coordinating disparate projects as part of a coherent whole;
- Centralizing authorization, security, and tracking systems;
- Unifying organizational solutions to managing copyright and IPR;
- Reducing duplication of effort and resources;
- Saving time for the creators and users through organizational structure and centralization of data.

Digital Asset Management Systems

DAM systems provide the means to manage digital assets from creation to publication and archiving. In some cases, systems can automatically take the data from the scanning, assign the image a network space, depending on the metadata that the creator assigns to it, and then store the digital object and metadata in a database. A DAM system may be as simple as a directory of files on a hard disk, each file containing a digital asset, with an accompanying database that stores descriptive and administrative metadata for each of the files. Each database record contains metadata that can be used to find and understand the asset, including the name of the file and probably information about its content, format, history and usage. A simple asset management system can be purpose-built from an off-the-shelf database management system, such as Filemaker Pro, Microsoft Access or one of the larger SQL database systems, like MySQL or Oracle.

Many cultural heritage institutions have computer systems that automate traditional functions. For instance, museums often have collection management systems for managing records of their artifact collections internally. Libraries usually have an online public access catalog for patrons searching the collection. DAM systems are a new addition to this family of computer systems. Once an institution has created a repository of reusable digital content (images of artifacts, video of public talks, etc), it will need to manage this new resource as a type of documentation collection, but one with special management needs. DAMs help the institution to manage these digital assets, and include information about the original artifact or event that the digital content relates to, as well as technical information about the digital resource, intellectual property rights to the digital resource (not the original artifact), and other types of relevant metatada that enable the institution to preserve and re-use these digital assets. Such a system can be used to administer a collection of assets, ensuring that each can be found and used by searching the data and locating the information about how to find the file. Usually, the system is intended to provide access for more than just simple management of the assets themselves, such as providing public access to the collection through a web site or using the assets to support the day-to-day activities of a museum or archive. However, delivery and asset management systems can be separate. It is also important to remember that a complete digital asset management strategy must start with the creation or acquisition of the assets. Tools that support the preparation of the data, particularly for processing batches of assets, are a very important part of the package.

In practice, such systems are almost always more complex than the simple one described above. There is usually more than one version of an asset to be tracked. For example, a primary image that is either digitized from a photograph or created directly with a digital camera may be processed manually to create a collection master; that master can then be mechanically manipulated to create derivative assets for specific purposes, such as a thumbnail and screen-sized version of the image for delivery on a web site. The result in this scenario is that there are four assets that all share an image content description, but each has its own administrative and technical description, and each may have different restrictions on its use. This example can be further complicated by considering a collection of digital images of objects or sites. There may be multiple photographs of the same object or site, each of which has a family of digital assets derived from it; the photographs share some description of the object or place and each photograph has some specific content description, all of which is inherited by each of the appropriate derivatives. While it is possible to exploit the relational database systems mentioned above to build a DAM system to handle more complicated situations, the technical expertise required is significant.

Unless the uniqueness of the project at hand requires a custom solution, it is probably better to use software already developed for digital asset management where possible, whether by purchasing a proprietary system or reusing a system developed at another institution. There are a number of proprietary products on the market that can be some part of a DAM strategy. Unfortunately, at this point there are no obvious products that stand out as a unified, off-the-shelf solution to the general problem, though there are products that are appropriate for specific kinds of collections. The best advice that can be given at this time is for an institution to analyze its priorities, collections and intended audiences in order to develop a checklist that can be used to evaluate the products available and choose among the ones that seem appropriate. Some DAM tools have also developed by the cultural heritage community itself; for instance, a tool has been developed at Berkeley which has been used by several museums in the MOAC project (http://www.bampfa.berkeley.edu/moac/imaging/index.html).

Library software companies and vendors that specialize in museum collections management systems are beginning to extend existing products and develop new ones that can be a part of a DAM strategy. These products tend to be large-scale, expensive systems, with vendors usually working closely with clients to develop complete “solutions.” In general, this path is probably most useful when considering DAM as a part of a complete collections management strategy. Terms to look for when looking at the systems include collections management, content management, and digital libraries.

Many more software products that are appropriate for DAM have been developed with large-scale web site management in mind. These systems are available for a wide range of prices, with widely varying functionality. Some systems are more specifically oriented towards text or image-intensive collections, while others can handle a wider variety of media. Terms to look for when searching for these systems include digital asset management, asset repositories, media catalogs, document management and content management systems. Note that the term “content management” is used very loosely, often to refer to software that is intended for administrative uses which would not be appropriate for DAM. However, note that some content management systems do include both DAM and delivery functionality (Emery 2002).

Developing a DAM strategy

Rarely is developing a digital asset strategy as simple as picking out which software package to buy and then implementing it with local constraints. It is unlikely that the needs of a project will be a perfect match for any software package; there will almost always need to be some customization to match the system with local requirements. It is either a matter of paying the vendor more for the customization of a system that does most of the job, or of buying or developing the additional tools needed. Unless the institutional context for the project includes strong technical support, a custom designed system will probably cost more than buying existing tools for the job. Even with good technical support, it pays to look closely before setting out to develop a system locally.

The key is to plan carefully, starting with the acquisition and creation of assets, looking at every process that leads to their ultimate uses and safe storage. Develop a checklist that addresses the five areas discussed below. What are the staffing implications that arise from each of these areas? What are the data inputs and outputs at each step? What are the access points to the data, both within the institution and to the public?

File management

The most basic DAM issue is the management of the files that contain or constitute the assets. These files must have unique names so that they can be unambiguously addressed. Some proprietary systems handle all of the file storage for the user. File names are assigned by the system and the whole file storage array is essentially opaque to the user. This certainly can be useful in that it frees the user from the worry of tracking each individual file and makes it easier to control access to the digital assets appropriately. The disadvantage can come when trying to recover from a disaster, when upgrading a system or moving from one system to another; the user is completely dependent upon the software for access to the files. Operating system utilities necessary to these processes may be unable to address the files. It is best to establish what restrictions a software package imposes on file management with the vendor before purchase.

In more open proprietary systems, and in custom systems constructed around databases, the issues of file management are a very important part of the DAM strategy that should be addressed thoroughly in the early stages of planning. Strategies for naming files that seem intuitively obvious early in the process may become a problem as the project scales up. Naming schemes should be systematic, exploiting directory names, file prefixes and suffixes. It may be helpful to exploit existing identifiers that are known to be unique, like an accession number for museum objects. For a collection of digital images, the accession number could be used as the file prefix (appending a “v1”,“v2”, etc., for multiple views of the same object), the file suffix could reflect the encoding type of the each image, and each type of derivative could be stored in a separate subdirectory.

Another major issue for file management is creating backup copies of the assets that can be reloaded in the event of mechanical failure or operator error. This is most often done using magnetic tape. Though a detailed treatment of this subject is beyond the scope of this *Guide*, some basic principles can be pointed out. Backup strategies are usually a combination of making copies of new or changed files frequently and making full copies of the file system periodically. Tapes should be reused in rotation, retaining certain copies for longer periods, in order to be able to ensure an acceptable risk of loss. For example a busy project may make daily backups of

files that change, with a weekly full dump of the file system. The daily tapes are then reused, with the weekly tapes being retained until a defined period has passed, at which point one of them can be retained for a longer period while the rest go back into the rotation. Many such schemes are possible and need to be investigated with respect to the specific needs of the project. Note that backup tapes also become assets that need to be managed. See Section XIV on Preservation for issues related to long-term retention of copies of digital assets.

Metadata definition and management

Metadata about digital assets are generally classified into three categories, all of which are useful in the management process: ***descriptive metadata***, which is about the content and form of the digital asset to enable search and retrieval; ***administrative metadata***, which is about the history of the asset and policies associated with it, typically information on creation, quality control, rights and preservation; and ***structural metadata***, which records information about the internal structure and relationship of resources to facilitate their navigation and presentation.

Any DAM effort must consider what metadata needs to be captured for digital assets early in the planning process. At this stage of technology (and for the foreseeable future) digital assets are pretty much unusable without metadata. The specifics of metadata for the different media are covered in other sections of the *Guide* and in the appendix on metadata, but there are some general principles that should be followed when planning a project.

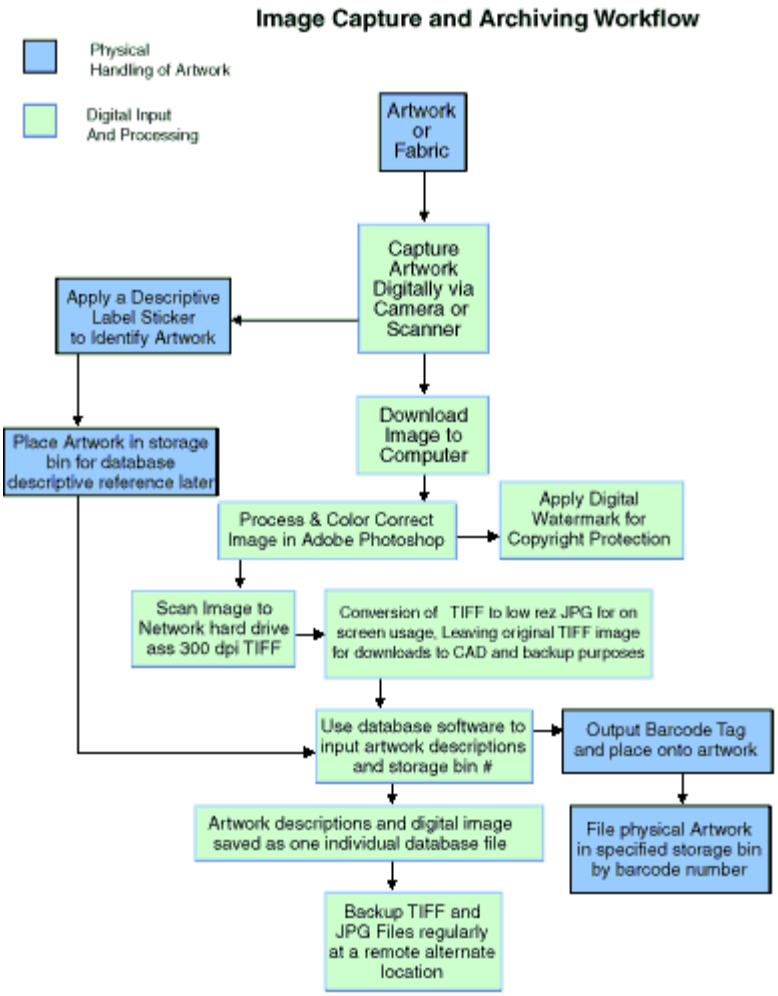
The effort associated with the creation of metadata can be equal to or greater than the effort to create the assets themselves, inasmuch as it represents an explicit statement of all the most important information about the asset. It is necessary to carefully consider the ways that the assets are intended to be used, who would be using them, and any legal or administrative requirements that are involved. Then the appropriate metadata must be captured or created to support all of the activities that require the assets, in addition to the file management activities described above and the workflow and access issues described below.

Because the DAM process is largely a function of managing metadata, careful attention should be paid to developing support for the creation process and for searching and reporting functions. If possible the system should be able to incorporate controlled lists of values (where appropriate) for metadata fields, either through user-defined lists or by incorporating standard lists, like thesauri and community-defined authority lists. The ability to search and to prepare reports from the metadata are both key functions in the DAM workflow.

Workflow

Any strategy for managing digital assets should address every step of the process, starting with the creation or acquisition of assets and metadata, through the processes of their day-to-day management, to delivery. Figure 1 shows the workflow for a process of capturing and archiving images, developed by Teri Ross (used by permission). This diagram illustrates that the DAM process can consist of many steps, requiring a variety of processes and possibly a number of different tools and skills. The development of a good digital asset management strategy should include an analysis of the necessary workflow in the planning stages of the project.

In any project the work of creating and acquiring both the assets and the metadata, and of migrating assets to new formats as necessary, can be substantial. Tools and utilities that make all of these processes more efficient and more reliable are very desirable. In particular, utilities that support batch processing of assets and the capture of metadata from the creation process are both very important parts of a DAM strategy. Note that many proprietary systems may not include all of the tools and utilities necessary to a good workflow so it is important to make sure that other software, purchased or created, can easily interface with the DAM system.



From Teri Ross, 'Digital Asset Management: The Art of Archiving', http://techexchange.com/thelibrary/DAM.html

Policy tracking and enforcement

A digital asset is only valuable to an institution if its use can be effectively managed. As is evident in other sections of this *Guide*, managing the rights associated with digital objects is complex but DAM systems can help resolve some of those complexities. Three areas where DAM systems can help manage rights include assertion, protection and management. Protection comes in various forms, from managing access to the digital repository, to tracking users, controlling what versions of material users can access, and ensuring that rights metadata is linked to the object when it is delivered to the user.

Data relevant to other types of policies may be a part of a DAM strategy. If the assets produce revenue for the asset holder, the DAM system must at least be able to interact with the fulfillment process. Versioning of assets is another area that can complicate a system. If new versions of the asset are created over time and all versions remain available to the user, it may be important to track the different versions and deliver a specific version on demand.

Policy tracking consists mostly of keeping the appropriate metadata about assets, and, in general, it is a fairly easy process to manage. The basic metadata management of a DAM system described above should be adequate. The systematic enforcement of policies through software is very much a research topic rather than a feature of systems that are available. It is reasonable to expect that systems should be integrated with the relatively simple access control that is available through the current generation of web servers. More sophisticated needs for policy enforcement should be discussed with vendors and/or developers with an eye to the long-term feasibility of the DAM strategy rather than as a necessary feature.

Access

In addition to their availability for public access, which is covered in the Distribution section of the *Guide*, digital assets are often needed in the day-to-day processes of the institution that owns them, possibly requiring the DAM system to interface with other systems. For example, a museum's exhibition design group may use images of artworks from the general collection in planning exhibitions. Their process would need to be able to access those images and their metadata.

In elaborate DAM systems, this kind of access is often part of the package. But even there, careful consideration should be given to the details of how the data is made available. Complete systems are built upon assumptions that may be overly restrictive, making it difficult to adapt to new demands in the future. In general, a DAM system should be able to accept input and deliver output in standard streams. The availability of good, open application program interfaces (APIs) with the system can ensure that interoperability can at least be developed.

Conclusion

Digital asset management at its most basic level uses technology, such as commercial off-the-shelf applications, to implement common sense goals of organizing resources, enabling users to discover them and owners to track them. At its most advanced level, a DAM system can radically transform the way an institution manages digitization and handles access to the digital assets. It is important to develop a complete strategy that covers the complete continuum from creation to delivery: storing the digital objects and their related data; supporting on-line searches; enabling users to retrieve the appropriate version of a particular digital resource depending upon their interest and/or access level; keeping track of rights restrictions; and limiting access to verified users. DAM makes workflow more efficient and can assist organizations save money and time. Digital objects are valuable assets for projects and institutions, and DAM systems offer a rich array of ways to ensure that these assets retain their value and are properly managed.

XIV. Preservation

Introduction

The projects and programs interviewed for this *Guide* all acknowledge that digital materials are at risk of loss, both through physical damage to the storage medium and through technological changes that render the data unreadable. Although none of the projects we interviewed identified digitization as a chosen method for preserving analog material, digitization is starting to supplant microfilm as a preservation medium and hence the urgency of providing long-term preservation of digital materials becomes of paramount importance. For museums and institutions with two- and three-dimensional objects, moving images, and audio, the need is even greater, since microfilm has never been even a minimally adequate preservation medium for these. Overall, there is an urgent need for significant research and the development of affordable strategies for trustworthy digital preservation services. This *Guide* does not aim to cover the full range of technical and scientific developments in storage media and methodology, but explores some of the issues involved in preservation in order to explain the practical steps you can take to store and preserve your digital materials.

In this *Guide* we define 'preservation' as the actions taken to preserve digital objects, whether they are created from analog collections or are 'born digital' objects such as original digital art, and digital audio or video that are not re-recorded from analog originals. This domain includes two important and overlapping areas: first, the goal of presenting the content of analog items via digital reformatting; and second, preserving digital content regardless of its original form. 'Conservation' of the original analog objects—generally the domain of conservators, museum curators, librarians, and collection managers—is somewhat different, as it involves the care and repair required to keep the physical object in existence. However, there is some overlap between the two: objects are often digitized to help conserve the original object by reducing wear and tear on fragile and vulnerable materials, while still allowing access to them.

Preservation of digital objects can be thought of as long term responsibility to the digital file — responsibility to ensure that the file retains the information it had when it was created. Keeping the content 'alive' over time entails much more than simply keeping the files intact. Issues such as metadata must be understood as well. These issues are explored more fully in the rest of the section. There are two distinct areas of intention here, although in practice they often merge: preserving the digital surrogate, so that its content remains accessible and usable over the long term; and using digitization as a method of preserving objects which are valuable, fragile or transient. For example:

- Access to digital surrogates replaces direct use of rare and fragile originals, aiding their conservation.
- Some projects use digitization to 'capture' objects that are discarded after the capture process, such as newspapers.

Therefore the digital surrogate is a form of preservation (although not a substitute for any other form of preservation) and must itself be preserved to ensure future access and use (see Section XI on sustainability). The project should consider the two areas: digital reformatting and preserving content in digital form.

Why preserve?

Digital preservation is an essential aspect of all digital projects: it is imperative that the digital objects created remain accessible for as long as possible both to intended users and the wider community. Decisions made throughout the digitization process will affect the long-term preservation of the digital content. While the processes involved in digital preservation are the same whether for digital surrogates or for born-digital objects, remember that born-digital objects are the only version of an artifact or object so are, in effect, the preservation copy. All these factors must be incorporated in your digital preservation strategies.

CEDARS (http://www.leeds.ac.uk/cedars/) has described the digital preservation process as follows:

> Digital preservation is a process by which digital data is preserved in digital form in order to ensure the usability, durability and intellectual integrity of the information contained therein. A more precise definition is: the storage, maintenance, and accessibility of a digital object over the long term, usually as a consequence of applying one or more digital preservation strategies. These strategies may include technology preservation, technology emulation or data migration. There is a growing wealth of information on digital preservation and related issues available on the Web.

The term ‘accessibility’ as used above deserves further glossing. It is important to distinguish between what we might term ‘machine accessibility’, involving the data integrity of the original files and digital objects, and human accessibility, or the ability to render the object intelligible and usable by a human reader. The latter is the more complex challenge, since it requires not only that all the original bits be present in the right order, but also that there be software available that can interpret them and display them appropriately.

Digitization projects are costly and time consuming, so it is essential that the digitization process should not need to be repeated, incurring more expense and subjecting the originals to further stress. The process may not even be repeatable: some projects we interviewed do not keep the original object. The Library of Congress, for example, converts newspapers to microfilm in certain cases, discarding them when the process is complete. The University of Michigan and Cornell have both discarded original objects for a variety of reasons, such as maximizing shelf space. Now the digital files are the *de facto* preservation copy, so it is even more important that they be suitable for long-term use and remain accessible for future generations.

Preservation should not be considered as an ‘add-on’ to the whole digitization process; the decisions taken throughout — choosing file formats, metadata, and storage — should all consider the long-term sustainability of the digital objects.

Preservation and storage of the digital surrogates

The projects surveyed saw digitization not as a way of creating preservation surrogates from original analog materials, but rather as a reformatting exercise designed to enhance access to and usability of materials. However, as noted above, projects like those at the University of Michigan, Cornell, or the Library of Congress, which discard the analog originals following digitization, must treat the digital version as a preservation surrogate. Furthermore, any project dealing with unstable materials must have similar concerns, since their digital version may in time become the only record of the original. Nonetheless, all of the projects interviewed recognized that they were creating assets that need to be preserved in their own right. Ensuring their long-term availability was particularly important where scholarly work depended upon or referenced them. As the limitations of microfilm become more apparent, many in the cultural heritage community are recognizing digitization as a means of preservation of digital content as well as improving access. Remember that here we are not so much concerned with conservation of the original but with preservation of the digital surrogate itself.

The starting point in planning digital preservation is to identify what can actually be preserved. Various levels of digital preservation are possible, from the ‘bucket of bits’ or bitstore (the storage of files in their original form with no plan to further reformat them for different forms of accessibility or searchability) through to preserving the full searchable functionality of an image or audio system. Each project must decide what is crucial, manageable and affordable. Consider a range of options, from the absolute minimum up to the ideal. If you decide to preserve the minimum, or the bucket of bits, then migration of software and operational systems will be less of a concern. This option might be attractive if your goal is simply to preserve archived data created by others, where your role is merely to keep the data, not to provide an ongoing framework for active use. However, if you decide to preserve the whole system you will need further migration of data to ensure usability in the future. The basic minimum will depend on each project, its rationale and its long-term user requirements. There is no limit to what can be preserved in digital format, apart from the limitations imposed by migration, storage space and other resources, such as staff time.

There are four main issues to consider in formulating your preservation strategy.

The first is software/hardware migration. All products of digital preservation must be migrated at some point, at the very least to a file format that the latest technology can recognize. If you have chosen to preserve the whole system, then operating systems and functional software must be migrated as well. This can cause problems as upward compatibility is notoriously uncertain, even from one version of software to the next, and there is no guarantee that the version you are using will be compatible with releases in many years to come. Full system migration must be carried out frequently to ensure access and usability. You will need to formulate a migration policy that is implemented on a regular basis rather than as a reaction to new software or hardware. Regular testing after migration is also crucial to ensure that functionality has been preserved. This is conventionally

called system emulation.

The second issue concerns the physical deterioration of digital media. All digital media deteriorate over time, but this process will be more rapid if they are stored in an inappropriate way, such as in a damp basement, or as a pile of CDs stacked one on top of another. Correct storage (e.g. in racks that enable the disks to be stored separately) and an environmentally controlled location will help to optimize their stability and protect them from loss. The Genealogical Society of Utah, for example, stores the archival digital master files in an underground vault in Utah as well as in another state in the USA. Tape storage is useful, as there is strong evidence for the extended lifetime of this medium when stored correctly; remember that all tapes must be spooled regularly to avoid sticking. As with software and hardware migration, digital media should be checked and refreshed regularly to ensure that the data are still readable, and this process should be part of your preservation policy. All institutions should draw up a policy that reflects the value of the digital material and therefore sets out how regularly the media should be checked and replaced. This aspect of the preservation policy is linked to hardware migration, as some media formats may no longer be readable by the latest computer hardware. Preserve your data on a medium where the hardware exists to transfer to later media if the original becomes obsolete. Remember that it is costly to use a data recovery agent to move files from an obsolete medium, so make sure your preservation policy will prevent this happening, and migrate while the process is still straightforward.

The third issue concerns metadata, which is crucial to preservation of digital resources. The level of metadata recorded and stored will, once again, depend upon what you choose to preserve, the minimum content or the fully functional system. The format of the metadata is also important, as the metadata should be as accessible as the data. ASCII will not need to be migrated, so it is the most durable format for metadata, although it lacks the functionality that a database might give. SGML/XML files (including HTML) are stored in ASCII format and provide a high level of structure and functionality without requiring proprietary software. The site structure and the relationships between the files should be recorded with a METS format XML file, and the other kinds of files (e.g. images, databases) that make up the site should, if possible, have preservation copies made. Word-processing files should be copied as ASCII text files, and if possible converted to XML to preserve their structure. All databases and Excel spreadsheets should be exported as comma-delimited or tab-delimited ASCII text files. PowerPoint files should be exported to HTML with gifs, with those gifs potentially re-saved as separate TIFFs. GIS datasets should be exported as comma-delimited ASCII.

A fourth and more complex issue is the question of user needs and preferences, which may cause certain formats to become effectively obsolete even while they remain technically functional. For instance, user discontent with microform readers threatens to make microfilm obsolete even though it still fulfills the original goals of its creation. User acceptance—and its decline—will be one of the key “trigger events” that will compel migrations to new delivery versions of digital collections.

Technical, media and file formats

Careful consideration of the technical issues will help you to ensure the long-term accessibility of your digital material.

As outlined above, the physical media on which the digital material is stored need to be managed in monitored, controlled systems and environments to avoid physical deterioration through human mishandling and inappropriate storage. Magnetic media such as digital audio and video tape are highly vulnerable to fluctuations in temperature and humidity, and of course to magnetic fields. Optical media such as CD-ROMs and DVD-ROMs are more durable, but should still be stored with care. For all media, it is advisable where possible to store a second full set in an off-site location, to guard against the risk of theft, fire, and similar disasters.

Choice of medium is equally important. Formats change rapidly and obsolescence is a perennial problem; it may be very difficult in the future to find a drive that can read ZIP disks, just as now it may be difficult to find readers for certain kinds of magnetic tapes. No one format can be guaranteed to persist and remain accessible in years to come, so you need to anticipate continuing developments and new products. Among currently available media, DVD, DLT (Digital Linear Tape) tape and CD-ROM are the most popular. Tape has a tried and tested longevity, while DVD and CD-ROM are newer media and current research has not yet agreed on their likely lifespan. Best practice in preservation is to migrate data from one medium to another, for example from optical disc to tape, while the hardware and software are still available. Migration is an integral part of any digital project and should be set out in your preservation policy.

If projects choose to make back-up or archive copies on fixed media, then it is crucial to ensure that the fixed media is stored correctly and that migration to the latest type (e.g. DVD) is carried out regularly.

A similar obsolescence problem will have to be addressed with the file formats and compression techniques you choose. Do not rely on proprietary file formats and compression techniques, which may not be supported in the future as the companies which produce them merge, go out of business or move on to new products. In the cultural heritage community, the *de facto* standard formats are uncompressed TIFF for images and PDF, ACSII (SGML/XML markup) and RTF for text. Migration to future versions of these formats is likely to be well-supported, since they are so widely used. Digital objects for preservation should not be stored in compressed or encrypted formats.

While TIFF and PDF are *de facto* standards, it must be remembered that they are proprietary formats with multiple 'flavors.' Not all image software can represent TIFF files accurately and Acrobat is required to view PDF files. Projects should monitor the changes in these formats and ensure that users will be able to access the files over time, either by migration or emulation.

Storage and retrieval

The readability of your files is dependent on application software and operating systems. Consideration of file format selection must take this into account. Access software, operating systems and user interfaces are subject to the same process of continuing development and change as storage media and file formats. New versions of operating systems will affect the usability of your files, even though the file format itself may still be accessible. Remember that your operating system may well be proprietary software; unless you are using a non-proprietary system such as Unix, you will have to purchase the more recent version and migrate the files.

It is advisable to separate the underlying content from the discovery and display systems.

Migration will be more complex for those projects that have chosen to preserve the complete integrated system, with full searchable functionality, as there are more features to capture and preserve. Backwards compatibility is often limited and so, as with media migration, these programs should be monitored closely to enable migration before obsolescence makes this impossible. For example, you may be using proprietary software which has regular (often yearly) upgrades. The versions may eventually change so much that backwards compatibility is no longer viable, resulting in, at best, a long and expensive process of data extraction and migration, and at worst, the loss of the data altogether. To ensure accessibility, migrate every file to every new version. The use of standard data formats such as SGML and XML can be an advantage here, since it keeps the data independent from the delivery software. Even if the latter becomes unsupported or if cost forces you to switch to a different delivery system, your data will remain intact and portable.

Link Box:

Preservation of Electronic Scholarly Journals: http://www.diglib.org/preserve/presjour.html

This project, a collaboration between CLIR, CNI and DLF working with the Andrew Mellon Foundation, aims to develop plans for affordable preservation strategies that will implement the OAIS reference model.

Monitor the user interfaces, too. Scripting languages, plug-ins, and applets may all become obsolete over time and users will not be able to access your materials.

In general, remember that standards are the key to creating a resource that is interoperable (the ability of the resource to be used by a different system) as well as sustainable over time. Relying on proprietary software or formats that are not accepted community standards may mean that you will have to migrate more frequently and in isolation. All projects will have to migrate their digital files at times, but by using standards this process will be less frequent and may take place when all projects are migrating data in response to a general change, such as that from SGML to XML, making it easier for everyone.

Organizational questions

The long-term view taken by institutions will also influence the sustainability of digital objects. Projects that have

established preservation policies and procedures, such as Cornell and the Online Archive of California, are identifying preservation issues and placing them at the heart of their institution's ongoing procedures. Insufficient institutional commitment to long-term preservation can create digital resources with limited sustainability.

Similarly, human and financial resources play a huge role in implementing such policies and procedures. Staff turnover and variations in usage will influence the level of commitment to individual projects. Unsatisfactory record keeping and metadata, administrative as well as preservation and access, can also contribute to institutional neglect of digital materials. Moving from individual projects to organizational digital programs will facilitate the long-term preservation of digital materials, by providing greater organizational stability and the overhead necessary to maintain good records and documentation.

Addressing the issues

Whether you have opted to preserve the minimum content or the whole discovery and display system, policies should be in place to ensure the long-term sustainability and accessibility of the digital files you have chosen to be preserved.

We found that attitudes to digital preservation vary enormously. Many of the major digitization projects are using research sources, such as D-Lib Magazine (http://www.dlib.org/), to help them understand the trends and strategies in digital preservation. At the same time, we found that others did not know what digital preservation was and were uncertain of its importance for their projects. There are many questions and no single answer to the preservation issue. No one project is satisfied that it has found the best storage and migration policy to ensure longevity. The Genealogical Society of Utah project is taking a larger view of the issues through its research with Brigham Young University to identify storage media and standards that will help ensure longevity of digital surrogates. All the projects agreed that adopting standards in creating digital objects would enable their sustainability. Each project will choose what it will preserve based on the level of resources available for storage and migration as well as user needs.

Research is being carried out that will help you make the right decisions for your digital resources. Projects, such as CEDARS, have attempted to establish strategies for ensuring digital preservation through technology preservation, technology emulation and data migration. VidiPax and other companies are addressing the technical media side through research into restoring and archiving digital media, while organizations such as NARA are addressing both traditional concerns (records management) and new technical standards. Internationally, The Digital Preservation Coalition in the UK is working to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally http://www.jisc.ac.uk/dner/preservation/prescoalition.html, and Preserving Access to Digital Information (PADI), a project of the National Library of Australia, maintains a useful web site with international research results and recommendations (see Link Box below).

The key to understanding and implementing digital preservation is to view it as a series of tasks rather than as one process. These tasks, if properly carried out, will help ensure not only that digital objects are correctly stored but also adequately maintained and useable over time.

Although there are others, six methods of digital preservation are widely discussed among heritage professionals. They are not mutually exclusive, and you may decide on one or a mixture of methods.

Technology preservation relies on preserving the technical environment in which the digital objects were created, both hardware and software. This approach is more relevant to projects that have chosen to preserve the full system with functionality and interface as well as the actual content. The digital files must be preserved as well as the computer and software. Maintaining hardware for long-term use is problematic, and it is also costly to store the physical machinery as well as the digital files. This method would apply to projects that wish to maintain the environment for historical use (a computer museum, for example) and may not be appropriate for many cultural heritage projects.

Technology emulation also seeks to preserve the technical environment, but instead emulates the environment on current technology, mimicking the software and even the hardware. Everything has to be emulated, including operating systems, scripts, and applications. This is unlikely to be a viable approach for most digital imaging projects. HATII's Digital Archaeology Report for the British Library (http://www.hatii.arts.gla.ac.uk/Projects/BrLibrary/index.html) has more detailed explanations of emulation.

Data migration focuses on maintaining the objects in current format and may involve migrating all files to a

newer format (for example in 2001, from JPEG to JPEG 2000), when current software no longer supports the earlier file format. Data migration is a viable option for cultural heritage programs and will involve making certain decisions, such as how often to migrate and what format to migrate to. It can be a time-consuming process, involving inevitable loss of data as well as sacrificing an element of the 'look and feel' of the original material. Using community standards at least gives the reassurance that every project will be migrating data of similar types at roughly the same point in time, making it more likely that tools and support will be available. The Archivo de Indias project has demonstrated that it is possible to migrate millions of documents form one file format to another, provided you design the work program effectively.

Enduring care encompasses all aspects of caring for the digital objects and the media they are stored on, including housing in secure locations and careful handling. This is short-term care and projects must consider migration and refreshing for longer-term preservation.

Refreshing is the process of moving data from one medium to another, e.g. from CD-R to DVD. It is not a total guard against obsolescence and should be seen as part of the whole digital preservation strategy.

'Digital Archaeology' describes what to do in a 'worst case' scenario, where a digital preservation policy has not been followed, or an unforeseen catastrophe has damaged the media. As described in the HATII report on Digital Archaeology, this is the process of rescuing content from damaged media or from obsolete or damaged hardware/software.

Storage Media

Data migration should not be confused with *'data refreshing'*. Data migration is the act of moving the actual content of the digital files to a newer file format (e.g. from Access 97 to Access 2000), while data refreshing is moving the files (no matter what format they are in) to a new physical medium, e.g. from CD to DLT tape, either because of obsolescence (e.g. laser discs) or because of physical deterioration.

Each project will have to identify the best storage strategy for its digital masters. This will include the choice of medium and the choice of institution/organization entrusted with managing that medium. The benefit of choosing a management service is that you can forego the responsibility and effort in choosing the medium. The benefit in investing the time to choose the medium is that you enjoy total control over your collection. Favored media vary from project to project and depend upon costs, access issues, physical storage space and technical equipment available. An in-depth examination of the technical issues of digital media is beyond the scope of this *Guide*. For a fuller explanation of these issues see the HATII report on Digital Archaeology cited above.

The formats used by the projects interviewed for this *Guide* include:

CD-ROM (including CD-R)

DAT Tape

DVD - both DVD - Video and DVD- ROM

DLT Tape

RAID Server

A mixture of the above media

Projects using audio-visual material will need to use a medium that can deal with the large file sizes created when digitizing such materials as well as enable access to the stream of data. DVD and DV Stream are the most popular media for these formats.

CD-ROM is a popular storage medium — it is cheap and simple to process. Projects such as the Online Archive of California, the Digital Imaging and Media Technology Initiative of the University of Illinois and the Colorado Digitization Project use CD to archive their digital materials. CD writers are easy either to install and run from the machine that is temporarily storing the data, or to back up from the server that holds the access data. CD-R has less storage capacity than DLT tapes or DVD — CD-R can hold approximately 640 MB of data while a DLT tape holds 35-70GB. Magnetic tape is known to be stable for approximately 30 years under good storage conditions. DLT tapes for instances should survive for a million head-passes over a 30 year period. Manufacturers claim that CD-Rs will last from 50 to 200 years, but experience indicates that even after as little as 10 years they are beginning to suffer from drop-outs. Media lifespans are an unknown quantity, so you may count on having to

migrate and refresh the data regularly. These figures are based on climate- and temperature-controlled storage conditions and safe environments. Lifespan also assumes that the media are free from manufacturing errors and are not mishandled.

You can act to improve the storage life of your media. We recommend that you handle media with care and that you check it regularly. The following two tables give examples of how to improve the lifespan of various storage media.

Good Practice Box:

Improving the lifespan of CDs and DVDs:

Avoid	Never	Always
Damage to the upper and lower surfaces and edges of the disc	Attach or fix anything to the surface of the CDs	Store media in a jewel case or protective sleeve when not in use
Scratching and contact with surfaces that might result in grease deposits (e.g. human hands)	Write on any part of the disk other than the plastic area of the spindle	If using sleeves, use those that are of low-lint and acid-free archival quality
Exposing discs to direct sunlight		Wear gloves when handling the master CDs

Good Practice Box:

Improving the lifespan of DLTs:

Avoid	Never	Always
Placing the tapes near magnetic fields	Stack the tapes horizontally	Keep tape in its protective case when not in use
Moving the tapes about	Put adhesive labels on the top, side or bottom of cartridge	Move tapes in their cases
	Touch the surface of the tape	Store the tapes in appropriate
	Put a tape that has been dropped in a drive without first visually inspecting it to make certain that the tape has not been dislodged or moved	Store the tapes vertically

There are two types of DVD: DVD-Video and DVD-ROM, which can be compared respectively to the audio CD and the CD-ROM. DVD-Video can contain a whole film with digital audio. The DVD-ROM stores computer-readable data.

Many larger projects store two copies of the media, in different storage facilities. Both the GSU and SHOAH use

underground vaults with temperature controls and high security in two geographically remote locations. SHOAH has the added risk factor of earthquakes and has chosen its second site in Washington DC. Not all projects can afford such facilities and the additional security that they provide, and will simply store one copy on site. We recommend, however, that all projects store archived master files off-site where possible to ensure against natural and human disasters.

We found that many projects did not check archived data for refreshing and did not have a preservation strategy at all, although many are developing them.

Stargazing Box:

There is a need for better long-term storage formats. The GSU project is conducting very interesting research into media for digital preservation. Through their work with BYU and Norsam (http://www.norsam.com/), they are examining media such as the HD-ROM, which has the following manufacturer's claims:

HD-ROM and the "ET" System are planned to have the following specification features: (ET is Electrical Transport Data Storage System)

- 200 Gigabytes per CD-sized disc
- 2 terabyte 10-disc Modular Reader
- 60 terabyte Modular Storage Unit which holds 30 10-disc Modular Readers
- Write rate of 30 megabytes per second
- Expandable to petabyte storage
- High-speed random access
- Each disc will have its own read head
- Read rate comparable to CDs

File formats

The file formats you choose are crucial to ensuring data migration. In 2001, uncompressed TIFF is the preferred raster image format for digital masters. Migration from this format may be inevitable to ensure long-term access to the materials. However, TIFF is widely supported and software manufacturers may continue to support it for the foreseeable future. It should be monitored carefully. TIFF is platform-independent. Archival copies should be stored uncompressed.

JPEG, JPEG 2000, MrSid, and GIF formats—which may be well suited for delivery—are not considered to be as well suited as TIFF for preserving digital master images, for two main reasons. All of these formats incorporate compression algorithms that offer potential benefits for storage and networked delivery, but increase the requirements for both monitoring and migration. Compression is one more thing to worry about when tracking the availability of software that can read and write formats without introducing information loss; thus, it is possible that intervals of migration will be more frequent for compressed formats rather than for uncompressed TIFF. In addition, because these formats employ lossy compression algorithms, there are opportunities for information loss at the time of capture and in subsequent migration activities. Other formats with lossless compression, such as PNG, are not widely used yet because of their relatively limited support within software packages and image viewers. TIFF, while not an open standard in the strict sense, is the *de facto* standard format for digital masters within the community, but not necessarily for delivery, given the fact that standard web browsers will not display TIFF images.

Formats for audio-visual material include WAV, AIFF, MPEG (various versions) as preservation archival copies. RealMedia (both audio and visual) and QuickTime are used for streaming data to users but not as archival formats.

Encoding for longevity

The structure of digital objects is just as important as media and format in ensuring their longevity.

Digital objects stored in a proprietary format, such as a proprietary database, are difficult to migrate forward;

proprietary software often has a short lifespan and is vulnerable to changes in the company that owns and supports it. A proprietary database can be unreadable within five years. You can minimize this risk by exporting your data in a standard format (such as tab-delimited ASCII). Similarly, use standard encoding schemes to structure your digital objects in order to minimize the risks of data obsolescence. Encoding standards such as SGML or XML are platform- and software-independent, and many projects, including some of the larger ones such as those at the Library of Congress, Cornell, Michigan, the Online Archive of California, and Virginia, have adopted these formats.

Metadata

One of the most important factors in ensuring digital preservation is the creation of robust metadata to describe the process of digitization as well as the actual digital objects themselves. Administrative metadata are particularly associated with digital preservation and the retrieval of the images from the archived media. This facilitates both short-term and long-term management and processing of digital collections and includes technical data on creation and quality control. Administrative metadata also handle issues such as rights management, access control and user requirements as well as preservation strategy information. The administrative metadata should include technical data that can be of potential use to systems (e.g., for machine processing tasks during migration), and to people, such as administrators monitoring the number and types of formats in the repository, or even to users inquiring about the provenance of a digital object. Having metadata in place will facilitate digital asset management, by helping to manage the digital files, their location and their content. Also see Section XIII on Digital Asset Management.

Link Box:

Some projects are adopting an institutional approach to metadata.

- Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/
- CEDARS project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
- Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metatada: http://www.nla.gov.au/padi/topics/32.html
- NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
- OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
- Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf
- Amy Friedlander, “The National Digital Information Infrastructure Preservation Program: Expectations, Realities, Choices and Progress to Date,” D-Lib Magazine (April 2002) Volume 8 Number 4; http://www.dlib.org/dlib/april02/friedlander/04friedlander.html
- OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
- A Recommendation for Preservation Description Information: A Report by The OCLC/RLG Working Group on Preservation Metadata: http://www.oclc.org/research/pmwg/pres_desc_info.pdf
- Anne Kenney et al., “Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism,” D-Lib Magazine (January 2002), Volume 8 Number 1; http://www.dlib.org/dlib/january02/kenney/01kenney.html

While administrative metadata are crucial to the long-term usability of digital materials, structural and descriptive metadata are also critical to the digital preservation process. All projects recognize the importance of good, robust metadata and have implemented schemes that facilitate the maintenance and accessibility of the resources. The importance of high quality metadata cannot be underestimated, as this can ensure that while data migration occurs, the original information about the digital object is not lost and can be recreated if necessary.

Organizational strategies

Over the last few years, those involved in developing and delivering digital objects have realized that preservation issues go far beyond questions of preserving bits and dealing with obsolete media and file formats. Indeed, there is widespread recognition of the need to develop trusted digital *repositories* that can ensure not only migration of media and formats, but also the preservation of relevant information about digital objects so that we can preserve their behaviors as they are migrated forward in time and make meaningful use of them in the future. We need to preserve not just the byte stream, but also the structure of the digital object and its context. Discussions about the attributes of digital repositories have crystallized around the emerging ISO standard, "Reference Model for an Open Archival Information System," originally developed by NASA to deal with the huge volume of digital material developed by the space program. (In May of 2002, RLG-OCLC issued a report based upon this model entitled "Trusted Digital Repositories: Attributes and Responsibilities.") The Draft standard states:

> An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.... The model provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives.

OAIS is a reference model, not an implementation plan. Libraries and archives are just now beginning to design digital repositories. [1] As they look to implement OAIS systems, metadata concerns loom large. Long-term preservation of digital objects in a repository requires that we develop and maintain structural and administrative as well as descriptive metadata about those objects. The METS initiative has gained much attention this past year as a flexible way to address these needs. The FEDORA project, which is complementary to METS, is aiming to develop behavior metadata as well. At this time, creators of digital objects and projects do not yet have full guidance on what digital repositories will need to secure their objects for the long term. In particular, much of the progress to date has focused on ingesting, storing, and preserving individual digital objects and much more research needs to be done on the long-term needs of complex digital creations. Nonetheless, attention to the emerging structure of repositories and their metadata needs will definitely increase the likelihood and ease with which digital objects can be incorporated into repositories and preserved for long-term access.

One of the major issues involved in creating trusted repositories is that of responsibility. To what extent will creators and local providers need to cede control over their digital objects to centralized repositories? The advantages that large, financially secure institutions can provide in terms of long-term preservation are unquestionable. Nonetheless, research efforts especially amongst the digital library community are also focused on developing distributive systems. Indeed, there is a growing sense that "lots of copies" will be an important preservation and authentication strategy for the foreseeable future. With the ever-declining cost of storage space, redundant distributive systems with shared responsibilities has many advantages that mitigate against centralized control.

Conclusion

Digital preservation is a discipline that is still in its infancy. Much of the preceding discussion provides a general description of various conceptual approaches to digital preservation, and also provides guidance for good custodianship (in the present) of existing digital files and digital media. However, it is expected that, in the near future, digital preservation will be executed within the context of enterprise-wide digital repository systems. This is an area of ongoing research and development. For example, in the United States, there is the new National Digital Information Infrastructure Preservation Program (Friedlander 2002), whereby the Library of Congress will lead a national planning effort for the long-term preservation of digital content and will work collaboratively with representatives of other federal, research, library, and business organizations.

Summary

Key factors in preservation are:

- Identifying the data to be preserved — basic content and metadata through to the fully functional program with interface look and feel
- Adopting standards for file formats
- Adopting standards for storage media
- Storing data on and off site in environmentally secure locations
- Migrating data
- Refreshing data
- Putting organizational policy and procedures in place

---

[1] Examples of projects using OAIS are CEDARS, NARA, NEDLIB and PANDORA. A chart for the deposit system for electronic publications can be found at: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html. The CEDARS project aims to produce strategic frameworks for digital collection management policies, and to promote methods appropriate for long-term preservation of different classes of digital resources, including the creation of appropriate metadata.

table of contents previous chapter

valid xhtml 1.1
abp~03/03

NINCH guide table of contents

Appendix A: Equipment

This appendix brings together material from various sections of the *Guide*, in expanded form, to provide a detailed account of the kinds of equipment that digitization projects typically require. This appendix provides a level of specificity which may not be essential to many project managers, but will indicate the level of detail needed to make certain kinds of purchasing and digitization decisions. Project managers need to understand the kinds of criteria that will be used for this kind of planning, even if they do not need to internalize the specifics.

Image and Video Equipment

Selecting the correct computer for digitization work is very important, but frequently overlooked. In particular image, video and audio processing is very processor intensive, as Ihrig and Ihrig point out, 'you never can have a fast enough computer or too much memory for scanning and processing images'. Therefore, at least one machine should be specified appropriately for this work. The key components to consider are memory, processor speed, the rate of data transfer between components, the size of the disk storage device, and audio and video cards.

When you plan to purchase a system you should test it to ensure that it meets your requirements. RAM-intensive operations such as image processing and video editing can take many seconds or even minutes in computers with dozens, rather than hundreds or even thousands of MB of memory. Design a test sequence to allow you to determine whether the machine will achieve your objectives.

Display devices

Do not neglect display where image work is concerned. Your computer display is likely to be the most used output device and a good quality display of sufficient size is important for effective and efficient quality control and comfortable working conditions. The display should ideally be 19" or 21" inch diagonal viewable measurement. Upgrading to a 21" display is a worthwhile investment. If the monitor is to be used for rigorous quality control ("soft proofing"), ensure that the display has hardware and/or software calibration, that the calibration mechanisms are compatible with the capture devices, and that standard color management profiles can be used. It is also important to control or calibrate the ambient environment along with the display device. If you are considering LCD flat panel monitors, they are available with analog or digital signal processors. The digital signal displays are better but not always compatible with all systems and video cards, so research the hardware carefully.

For many digital imaging projects, the resolution of the monitor can be far more important than the size of the screen. The monitor or screen resolution is the density of phosphor dots on the Cathode Ray Tube (CRT) or pixels in the matrix of a LCD monitor. This is usually measured in dot-pitch — e.g. 0.28mm dot-pitch is approximately 90 dots per inch (dpi).

Calculating Screen Resolution from Dot-Pitch:

You may find it useful to use dot-pitch numbers provided by computer suppliers to work out display dpi.

Screen Resolution = (1/dots per mm) x mm per inch

so a screen claiming a dot-pitch of 0.28 mm would have a resolution of 90dpi

90 = (1/0.28) x 25.4

Explanation —

0.28 mm ‘pitch’ means 1/0.28 dots per mm

1/0.28 dots per mm is equivalent to 3.5714 dots/mm

(constant 25.4 mm in an inch)

3.5714 * 25.4 =approx. 90 dots/inch

Relationship between Image dpi, Screen dpi and Magnification:

Level of Magnification = image dpi/screen dpi. Thus, regardless of the physical screen dimension (15” or 21” monitor), a 300 dpi image displayed on a 75 dpi monitor will be magnified 4:1.

Note that screen resolution can be measured in pixel dimensions, as well as dpi.

Pixel dimensions (both of image and monitor) are important if you are concerned about scrolling. (For example, if a monitor is set to a 1024 x 768 setting, regardless of the dpi of the image, if the image has more than 1024 pixels in its width or more than 768 pixels in its height, one will need to scroll in one or both directions.)

Relationship between Pixel Dimension Settings (controlled by software, not fixed by hardware) and “screen real estate”:

Setting	Screen Real Estate	Gain
800 x 600	480,000 pixels	
1024 x 768	786,432 pixels	+63.8% over 800 x 600
1280 x 1024	1,310,720 pixels	+66.7% over 1024 x 768; +273% over 800 x 600
1600 x 1200	1,920,000 pixels	+46.5% over 1280 x 1024; +244% over 1024 x 768; +400% over 800 x 600

The lower the dot-pitch is, the higher the resolution of the monitor. Although the quality of flat panel LCD displays is improving and they are becoming increasingly affordable, in 2002 they do not yet achieve the same price-to-quality ratio of conventional CRT monitors. A final consideration in relation to display is the video adapter, or video card; often now built into the motherboard of computers you should expect to see at least 16MB of video RAM or more, as less than 16MB RAM will cause problems when viewing and manipulating large images.

Printers

When buying a printer, projects will need to consider the dimensions of the original material, the desired dimensions of the output, and whether color is required, as well as the printer resolution. Printer resolution is a measure of the printer’s ability to produce individual dots, and is usually expressed as dots per inch (dpi). Printer resolution may also be expressed as lines per inch (lpi).

Color ink jet printers and associated media (inks, papers) have become very affordable and quality is often fit-for

purpose. The variables of paper, ink, storage environment, and use all determine the life expectancy of their output. If one of the planned outcomes for the digitization project, however, is the ability to create the equivalent of photo prints directly from the digital source, it is best to consult the photography department at another institution or a reputable printing service bureau or advice on printer selection.

Digital imaging devices

Definition Box:

Capture Device	Details	Scanning Array
Digital Camera	Linear Sensor (tends to be slow)	Linear CCD Array
Digital Camera	1 shot cameras with color diodes	CCD Area Array
	3 shot cameras with color filters	CCD Area Array
Flatbed Scanner	Medium to high-end devices, better optical image quality	Linear CCD Array
	Low-end devices, poorer optical image quality. Rely on interpolation (software algorithms) to achieve their resolution	Linear CIS/CCD Sensor
Engineering Scanner	Media pulled past the scanning array rather than the array moving over the media. Typically range from 36” to 62” width capacity by up to _” think capacity.	Linear CCD Array
Drum Scanner	Typically high-end prepress or graphic house equipment. Needs specialist operator. Media tightly taped to a drum with a layer of oil and rotated round the PMT at high speed. Color correction of negatives can be problematic.	PMT
Hybrid Drum Scanner	Halfway between a drum scanner and a flatbed. Can handle a wider variety of media than transparency scanners. Media is held tight over a cylinder by a set of rollers. This has the advantages of a drum scanner, but without tape and oil.	CCD
Transparency Scanner	High resolutions suitable for 35mm negatives and slides. Higher end machines have more advanced software for accurate color correction of negatives.	CCD Area Array
Microfilm Scanner	Available for 16mm, 35mm and microfiche formats. Usually output in bi-tonal or grayscale. “Blipping”, or the ability to distinguish individual frames, is the key to high throughput and accurate results.	Linear CCD Array
Book	Can be overhead (planetary) non-contact scanners or face down	Linear CCD

Scanner	(contact) cradle scanners. Also suitable for large documents.	Array
Card Scanner	Designed for business card, check or fingerprint card scanning. Can be suitable for card catalogs.	Linear CCD Array

Audio Equipment

Institutions may find themselves with a rich array of materials in analog form, but without the devices to play this material back. Unlike textual and still image material (with the exception of slides and born digital), audio and moving image material require a playback device in addition to a digital capture device. For example, a flatbed scanner can digitize directly a wide range of reflective media of different formats and sizes (e.g., photographs, letters, printed matter, bus tickets). No similar general-purpose capture device for audio and moving image material exists. A collection that included 78 rpm records, compact cassettes, 8mm film and VHS video cassettes would require a playback device for each of these and each would then need to be connected to an appropriate digital capture device. For audio and moving image material that is already in a digital format (such as CD or Digibeta), playback equipment is less of a problem. Although many--frequently incompatible--proprietary digital formats exist, their recent development means suitable playback equipment is still on the market and relatively easy to source. Therefore this section concentrates on identifying analog audio and moving image formats, their properties and the source device required.

Three methods can be used to progress from moving image film to digital. Film can be transferred onto videotape for digitization via a transfer box or multiplexer. Both of these options depend upon the material being projected in some way. Transfer boxes project the image into a box containing a mirror and onto a rear image projection screen with the video camera mounted on the other side of the box. The resulting video is subsequently digitized. These transfer boxes are not expensive, but do not in general produce quality as high as a multiplexer because they introduce generational loss.

A better solution is to use a multiplexer. In this device the projector and camera are mounted on a single table. The image is projected by a set of lens and mirrors, directly into the camera without the need for a projection screen. This has advantages for image clarity. In both processes quality suffers because it introduces an extra production generation into the reformatting of the analog material. An alternative to these methods is the use of 8, 16 and 35mm film for a chain film scanner to digitize directly from the analog film material. These machines scan the films and digitize at the scanner, passing the digital signal to the computer. (They work slightly differently for digital video. In this instance they grab individual lines of video to construct a frame and produce broadcast quality digital video.) In 2001 the costs of these machines remains high at between $500,000 and $1,000,000. One of the strengths of chain scanning is that, because the analog to digital conversion is done at the camera rather than on the computer, there is less opportunity for noise to be added by the process to the analog signal. Whereas small institutions can probably set up a transfer box or multiplexer system, even wealthy institutions would find outsourcing to a facilities house to be the only practical option if they wished to go directly from the analog film to the digital material.

Determining a film's original frame rate is also difficult without viewing the film with a projector, particularly for old 8 and 16mm films. The widespread availability of VHS and S-VHS video players makes the playback of these video formats for digitization relatively simple. The rapid adoption of digital formats in broadcasting, post-production and amateur markets is making the availability of even quite recent analog video devices scarce.

As there are fewer analog audio formats, these provide less of a problem than moving images. Compact cassette players, 33 and 45 rpm record players are still widely available new. Even record players with a 78 rpm speed can still be purchased new. The other formats present a greater challenge. If digitizing the sound as played on period equipment is important the tone arms of phonographs and gramophones can be customized to provide an appropriate feed. Alternatively, the sound can be recorded via an external microphone onto a more convenient intermediary format. Reel to reel tape, wire recorders and cartridges pose similar problems of transfer. By modifying the equipment, it may be possible to provide a direct sound output. Alternatively, the sound can again be captured via an external microphone to an appropriate intermediate format. Here is where a great deal of specialist advice can be helpful. Just as we noted that it is easier to train a good photographer in digitization than it is to train a digital expert in photographic principles and practice, you will find that sound engineers bring to the digital environment strengths that are difficult to replicate.

In the case of all audio and moving image material, whether it is in analog or digital form, projects should carefully consider the advantages of outsourcing digitization. In general audio and moving image digitization require more and more expensive and specialized equipment than is necessary for still image material.

Audio Media	Properties	Source Device
Wax or Celluloid Cylinders	1890s & 1900s, up to 5”diameter, 2-4 mins. playing time	Phonograph. See http://www.tinfoil.com for details of digital transfer.
Wire	Magnetic coated wire drums or reels. Invented 1898. Widely used by the US military in WWII. Eclipsed by magnetic tape by the mid 1950s.	Wire Recorder
78 rpm shellac resin discs	1898 to late 1950s, 10”(25cm) and 12”(30cm) most common sizes	Gramophone (wind-up) or Hi-Fi. Gramophone’s steel needles need replacing after each side or record played. Hi-Fi needs a 78 rpm turntable and a cartridge with a 78 rpm stylus. For best results on modern equipment a phono pre-amplifier is required to correctly equalize the different types of record.
45 rpm and 33 rpm vinyl discs	7” (20cm) single and 12” (30cm) long play. Long play (LPs) introduced in 1948, stereo recordings in 1958.	Hi-Fi. Hi-Fi requires turntable with 45 and 33 rpm speeds.
Reel to Reel magnetic tape	½” to ¼” magnetic tape. BASF and AEG developed 6.5mm ferric tape and Magnetophone player in Germany from 1935. Post-war development in USA by Ampex and 3M. Stereo capability from 1949.	Reel to Reel player for appropriate width of tape.
Compact Cassette	Magnetic polyester tape introduced by Philips in 1963.	Hi-Fi. Hi-Fi requires compact cassette player.
Cartridge	¼” magnetic tape. Fidelipac (4-track, devised 1956, released 1962) and Lear (8-track, 1965) cartridge systems.	Despite similarities 4 and 8 track cartridges are not compatible and require separate players. Predominantly used for in-car audio. 4 track unpopular outside of California and Florida.

Storage devices and systems

File storage requirements are a major consideration with any digitization project, and particularly for projects dealing with digital images or video. While storage devices have grown steadily since their invention and show no signs of reaching a plateau, storage demands have kept pace with them, and file storage remains a concern for delivery and for backup. Consider that 150 24-bit color TIFF files at 400 pixels per inch (ppi), plus 150 JPEG files at 100 ppi, together with associated metadata, would occupy 5 - 6 GB. In 2002, this is approximately the

entire capacity of most standard PC internal hard drives.

Different types of analog material (text, images, audio) and different ways of encoding it will have a direct impact on the sizes of the files and the quantities of storage that a project will require. A terabyte, or 1000 gigabytes, offers sufficient storage capacity to hold the equivalent of 300,000,000 pages of ASCII text, 20,000,000 pages of bitonal scanned documents, 1,000,000 pages of color images, 1,800 hours of recorded sound, or 500 hours of good quality video.

Projects have various storage options, of which the most common are internal hard drives, optical drives, tape drives and networked hard drives. Internal computer hard drives are fairly fast and capacious, but since they receive daily use they are at risk for hardware failure, infection by viruses, and similar calamities. Since they do not use a removable medium, they need to be backed up onto some other storage format regularly. Optical drives (including CD-R and DVD-R) are removable and fairly durable, although their capacity is limited (640 Mb for CD-R, 6 Gb for DVD-R). Tape drives provide much greater capacity, but their access speed is comparatively slow and the tapes require care and protection to avoid loss of data. There are at least two types of digital tape: DAT (digital audio tape) is the slower and more unreliable of the two, and offers a somewhat lower capacity (on the order of 12-20 Gb without compression), but uses cheaper equipment. DLT (digital linear tape) is faster, more reliable, and offers capacity in the range of 35-70 Gb without compression, but the equipment is more expensive. A newer format, AIT (Advanced Intelligent Tape) offers even greater capacity but slow speeds. Networked hard drives offer increased speed and capacity; in the case of a RAID array (which uses multiple drives acting in tandem) there is also the possibility of complete data redundancy and protection against loss. These are also among the most expensive options when expressed as cost per megabyte of storage.

Another distinction worth examining is the difference between online and offline storage devices, each of which has its own advantages and special considerations. Online storage (i.e. storage which, like a disk drive, is always attached to a live computer), provides immediate “random access” to any part of the storage system. Offline storage (or removable storage, such as tapes or CD-R disks) typically provides a lower cost per Mb of storage, but has the disadvantage that when you want to access a file you must locate the media on which it is stored, fetch the media, and load it. It is also typically slower in actual access speed (the speed with which data can be read from the disk or tape) than online storage. Hierarchical storage systems (or Storage Management Systems — SMS) combine elements of online and offline storage along with special control software to provide seamless access to a combined system.

Each of these options is ideal for some purposes and not for others, so when purchasing storage you should consider where your needs chiefly lie. For backup and preservation purposes, it is essential that the storage medium be removable and fairly durable; depending on how much data you need to back up, it may also need to be high-capacity. For smaller projects, or projects dealing with smaller files (such as text files in SGML/XML), backing up onto CD-ROM or DVD may be sufficient. (Remember that you should regularly back up not only your actual digital assets, but also all of your work files, including administrative materials, tracking records, correspondence, and so forth; be sure to include these in your estimates of required backup capacity.) For high-capacity backup, such as may be required by large image or video projects, tape storage may be necessary. The requirements for preservation are very similar, with the proviso that reliability and long-term durability may be even more important factors to consider. Most of the projects surveyed for this guide use either CDs or tape for external backup.

For publication purposes (i.e. for storing the data that is being actively delivered online) speed is essential; no removable storage devices are sufficiently fast. For large projects with substantial amounts of data and high demand, a high-speed, high-capacity system such as a RAID will be required. A RAID (Redundant Array of Independent Disks) is a collection of drives acting in tandem, which offers not only very high read-write speed, but also the option of full data redundancy (if one drive fails, the system can “fail over” to another drive seamlessly, and the failed drive can be replaced without interrupting data access). Such systems are very expensive, but if your data is mission-critical and you cannot afford any down time, their reliability is worth the investment. It may also be worth considering outsourcing storage of this type, or working with a consortium to share storage costs among several institutions.

Projects may also wish to consider data warehousing, a data repository specifically structured for querying and reporting. Data warehousing is a process used by large businesses to ensure sustainability and access to data about the company but cultural and heritage institutions may find the process or even just the concept, to be useful in storage planning. Several variables are worth mentioning here:

- Subject-oriented – Data that gives information about a particular subject instead of about a company’s ongoing operations.

- Integrated – Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- Time-variant – All data in the warehouse is identified with a particular time period.
- Non-volatile – Data is stable in a data warehouse. More data is added, but data is never removed. This enables management to gain a consistent picture of the business.

NINCH guide table of contents

valid xhtml 1.1
abp~03/03

NINCH guide table of contents

Appendix B: Metadata

This appendix brings together material from various sections of the *Guide*, in expanded form, to provide a detailed description of the kinds of metadata and metadata standards that are of greatest importance to the cultural heritage sector. It has already been observed, but is worth repeating, that metadata is a crucial part of any cultural heritage digitization project and should not be neglected. Without metadata identifying its source, contents, and details of creation at an absolute minimum, a digital object is useless. Capturing additional information to facilitate rights management, administrative tracking, preservation, and distribution can enable you to get much more powerful use from your digital materials with much less difficulty.

What is metadata?

Metadata is literally "information about the data": information created about the source material and the digital version to record the essentials of their identity, creation, use, and structure. Their purpose is to facilitate the discovery, use, management and reusability of digital material. Metadata can be usefully divided into three categories: descriptive, administrative and structural. These are not rigidly bounded groups and they frequently overlap.

Descriptive metadata describes and identifies information resources, to facilitate searching, retrieval, and management. It typically includes basic bibliographic information such as the creator, title, creation date; catalog information such as accession or other identifying numbers; and topic information such as keywording. Examples of descriptive metadata include Library of Congress Subject Headings, Categories for the Description of Works of Art (CDWA) (http://www.getty.edu/research/institute/standards/cdwa/), the Art and Architecture Thesaurus (AAT) (http://www.getty.edu/research/tools/vocabulary/aat/), and the Dublin Core metadata set.

Administrative metadata is used to facilitate management, tracking, migration and re-use of digital assets. It typically includes information on creation, quality control, rights and preservation. See Cornell University's "Web Hub for Developing Administrative Metadata for Electronic Resource Management" (http://www.library.cornell.edu/cts/elicensestudy/home.html). The term "technical metadata" is also used in a similar sense to indicate metadata about data capture and the technical characteristics of the images.

Structural metadata describes the internal structure of digital resources and the relationships between their parts. It is used to enable navigation and presentation. Examples of structural metadata are included in the METS standard (http://www.loc.gov/standards/mets/) and SMIL (Synchronized Multimedia Integration Language) (http://www.w3.org/TR/REC-smil/)

Text metadata

Metadata for textual resources is in some ways the most straightforward to create, because it can be captured in the same format as the digital object itself, and can be included directly in the digital file, for instance as a header section in an SGML/XML-encoded document. EAD, TEI, and HTML all include a header element of varying scope; the HTML header provides for the inclusion of basic Dublin Core metadata terms, while the EAD and TEI headers provide for more extensive information about both the electronic file and the information it captures or describes.

Image metadata

Metadata for still images may be stored in a file header or in a separate database or file. Because images themselves, unlike text, cannot currently be searched very effectively for content, metadata is doubly important for retrieval purposes, as well as for internal project tracking, management, and documentation. The METS

standard can be used to bring together the different types of image metadata required for different project purposes. It can also be used not only to document individual images, but also to represent the relationships between multiple images that together constitute a single digital object (for instance, high-resolution archival images, thumbnails and delivery images at lower resolutions, images of particular details at higher magnification). The NISO IMG draft standard on metadata requirements for digital still images provides extremely detailed specifications for capturing technical metadata for images: http://www.niso.org/standards/dsftu.html and http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

Audio-visual metadata

As with still images, metadata is crucial to digital audio or video, and the principles of metadata interoperability and documentation standards are as important to digital AV media as to still image and text media. Metadata for digital audio and visual resources can be used in much the same way as metadata for complex digital objects composed of still images. A metadata standard like METS (with the appropriate extension schema) can be used to describe the structure of an audio-visual digital object: for instance, a group of original taped interviews and a final version edited for delivery.

SMIL (Synchronized Multimedia Integration Language) can be used to describe the content and structure of time-based digital files such as audio and video. SMIL, for instance, can be used to describe structural metadata about a particular frame of video (frame 30, timecode 01:20:36.01) as well as to link the appropriate series of frames to alternate representations such as a transcription of the dialogue in that scene. As with image resources, this allows users to search for a particular bit of dialogue or the name of a character, and be taken directly to the video scene in which they appear.

Link Box:

Key AV Metadata Sites

- Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
- Synchronized Multimedia Integration Language (SMIL): http://www.w3.org/AudioVideo/
- Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
- MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
- Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
- Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
- Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
- Library of Congress's Digital Audio-Visual Extensions to METS Standard: http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
- Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html

Metadata standards

A number of metadata standards are now in use by the cultural heritage community that have been developed by different subcommunities to address particular needs. These standards are not mutually exclusive; on the contrary, some of them, such as METS, are specifically intended to be a way of bringing together various forms of metadata in a single place where it can be processed uniformly and predictably.

METS

The Metadata Encoding and Transmission Standard (METS) is an XML-based encoding standard for digital library metadata. It is both powerful and inclusive, and makes provision for encoding structural, descriptive, and administrative metadata. It is designed not to supersede existing metadata systems such as Dublin Core or the

TEI Header, but rather to provide a way of referencing them and including them in the METS document. As a result, it is an extremely versatile way of bringing together a wide range of metadata about a given digital object. Through its structural metadata section, it allows you to express the relationships between multiple representations of the digital object (for instance, encoded TEI files, scanned page images, and audio recordings), as well as relationships between multiple parts of a single digital representation (for instance, the sections of an encoded book). Its administrative metadata section supports the encoding of the kinds of information projects require to manage and track digital objects and their delivery: technical information such as file format and creation; rights metadata such as copyright and licensing information; information about the analog source; and information on the provenance and revision history of the digital objects, including any data migration or transformations which have been performed. METS is a very recently developed standard but is well worth watching and using.

Dublin Core

The Dublin Core Metadata Element Set defines a set of 15 essential metadata components (for instance, author, title, format) which are broadly useful across disciplines and projects for resource discovery and retrieval. These components can be used to add metadata to HTML files (using the <meta> tag) but can also be used in other contexts to create basic metadata for a wide range of digital resources. Dublin Core does not provide for detailed administrative or technical metadata, and as such is largely suited for exposing resources for search and retrieval, rather than for internal resource management and tracking. In addition, since its goal is to be simple and broadly applicable to a wide variety of resources, it does not provide for the kind of highly structured metadata about specific document types that TEI and EAD offer. Although projects using these encoding systems will probably not need to use the Dublin Core, they may find it useful to be aware of it as a possible output format for distributing metadata about their resources. One aspect of the work of the Consortium for the Interchange of Museum Information (CIMI) is research into SGML, XML, and metadata standards such as Dublin Core for museum collections and RDF.

TEI Header

The TEI Header is a required component of any file conforming to the Text Encoding Initiative Guidelines, and is ordinarily used to document a text file encoded in TEI. However, it can also be used to describe other kinds of resources. It is designed to express a wide range of metadata about a digital file, whether that file is an encoded text, an image, a digital recording, or a group of any of these. It provides not only for standard bibliographic information about the file itself and about its source, but also more specialized metadata to record the details of classification schemes, encoding and sampling systems used, linguistic details, editorial methods, and administrative metadata such as the revision history of the file. It is designed to accommodate a wide range of metadata practices, and while it offers highly structured options for capturing detailed metadata, it also allows for briefer and more loosely organized headers which record only the most basic information.

EAD

In a sense, the Encoded Archival Description (EAD) bridges the realms of data and metadata. As a digital finding aid, it may stand on its own as metadata about an archival collection. As a digital representation of an analog finding aid, it may also be a form of digital preservation (particularly if the original finding aid has any historical significance). It provides for the capture of all the information ordinarily conveyed in a finding aid, but it also provides for metadata about the finding aid itself—its author, language, publication details—and about the EAD file as well. EAD is a powerful tool for providing digital access to archival collections by representing the information user's need to discover archival materials of interest in a consistent and digitally transparent way.

SPECTRUM

The UK Museum Documentation Standard represents a common understanding of good practice for museum documentation, established in partnership with the museum community. It contains procedures for documenting objects and the processes they undergo, as well as identifying and describing the information which needs to be recorded to support the procedures. Spectrum was developed by the MDA in Great Britain. CIMI has adapted the SPECTRUM XML DTD for web based museum object.

NINCH guide table of contents

valid xhtml 1.0 strict
abp~03/03

NINCH guide table of contents

Appendix C: Digital Data Capture

This appendix brings together material from various sections of the *Guide*, in expanded form, to provide a detailed description of how analog information is converted into digital data in various media types. While for many purposes this level of technical detail may be more than is needed, a basic understanding of the principles involved can be useful in evaluating the appropriateness of certain types of equipment, determining when digitization is likely to yield good results (or not), and understanding why certain kinds of conversion can result in data loss or degradation. Specific recommendations for formats, settings, and how to get the best results from different kinds of materials are addressed in the main sections of the *Guide*; the goal here is to provide a more detailed explanation of the basic principles involved.

General Principles

Analog and digital data are fundamentally different: where analog information is generally smooth and continuous, digital information consists of discrete chunks, and where analog information bears a direct and non-arbitrary relationship to what it represents, digital information is captured using formal codes that have only an arbitrary and indirect relationship to the source. Thus while an analog image, for instance, consists of continuously varying colors and shading, a digital image consists of a set of individual dots or pixels, each recording the color intensity and other information at a given point. Although the specific kinds of information vary from medium to medium—sound waves, light intensities, colors—this basic difference remains a constant.

Conversion from analog to digital thus requires that the continuous analog information be sampled and measured, and then recorded in digital format. There are several basic factors which govern this process and which determine the quality of the digital result.

The first of these is the density of data being captured from the analog original: in effect, how often the original is sampled per unit of time (in the case of video and audio) or area (in the case of images and video). For digital audio, the higher the sampling rate, the smoother the transitions between the individual packets of sound, to the point where, with modern digital audio, they cannot be detected by the human ear. A low sampling rate results in clipping, the audio equivalent of jerky animation. For digital images, the higher the sampling rate (i.e. resolution), the smoother and less pixellated the image appears, and the more it can be magnified before its granularity becomes visible.

The second factor at work is the amount of information that is recorded in each sample. Individual pixels in an image may contain very little information—at the most minimal, they may take only one binary digit to express on versus off, black and white—or they may take 32 bits to express millions of possible colors. Large sample size may be used, as in digital images, to capture nuance, finer shadings of difference between values. They may also be used to express a wider total range, as in the case of digital audio, where a higher frequency response means that the recording can capture a greater range of frequencies, with higher highs and lower lows.

Both sampling frequency (or resolution) and sample size (frequency response, bit-depth) involve a trade-off of data quality and file size. It is clear that the more frequently you sample, and the more information you capture in each sample, the larger your file size will be, and the more costly to create, transmit, store, and preserve. Decisions about digital data capture are thus not simply a matter of achieving the highest possible quality, but rather of determining the quality level that will represent the original adequately, given your needs. Various sections of the *Guide* explore these considerations in more depth.

The remainder of this appendix describes how these principles apply in detail in particular digital media.

Digital Audio and Video Capture

In analog audio recording, a plucked string (for example) vibrates the air around it. These airwaves in turn

vibrate a small membrane in a microphone and the membrane translates those vibrations into fluctuating electronic voltages. During recording to tape, these voltages charge magnetic particles on the tape, which when played back will duplicate the original voltages, and hence the original sound. Recording moving images works similarly, except that instead of air vibrating a membrane, fluctuating light strikes an electronic receptor that changes those fluctuations into voltages.

Sound pressure waveforms and other analog signals vary continuously; they change from instant to instant, and as they change between two values, they go through all the values in between. Analog recordings represent real world sounds and images that have been translated into continually changing electronic voltages. Digital recording converts the analog wave into a stream of numbers and records the numbers instead of the wave. The conversion to digital is achieved using a device called an analog-to-digital converter (ADC). To play back the music, the stream of numbers is converted back to an analog wave by a digital-to-analog converter (DAC). The result is a recording with very high fidelity (very high similarity between the original signal and the reproduced signal) and perfect reproduction (the recording sounds the same every single time you play it no matter how many times you play it).

When a sound wave is sampled using an analog-to-digital converter, two variables must be controlled. The first is the sampling rate, which controls how many samples of sound are taken per second. The second is the sampling precision, which controls how many different gradations (quantization levels) are possible when taking the sample. The fidelity of the reproduced wave can never be as accurate as the analog original; the difference between the analog signal and the closet sample value is known as quantization error. This error is reduced by increasing both the sampling rate and the sampling precision. As the sampling rate and quantization levels increase, so does perceived sound quality.

In digital representation, the same varying voltages are sampled or measured at a specific rate, (e.g. 48,000 times a second or 48 kHz). The sample value is a number equal to the signal amplitude at the sampling instant. The frequency response of the digital audio file is slightly less than half the sampling rate (Nyquist Theorem). Because of sampling, a digital signal is segmented into steps that define the overall frequency response of the signal. A signal sampled at 48 kHz has a wider frequency response than one sampled at 44.1 kHz. These samples are represented by bits (0's and 1's) that can be processed and recorded. The more bits a sample contains, the better the picture or sound quality (e.g., 10-bit is better than 8-bit). A good digital signal will have a high number of samples (e.g., a high sampling rate) and a high number of bits (quantizing). Digital to digital processing is lossless and produces perfect copies or clones, because the digital information can be copied with complete exactness, unlike analog voltages. High bit-depth is also result in much-increased dynamic range and lower quantization noise.

Ideally, each sampled amplitude value must exactly equal the true signal amplitude at the sampling instant. ADCs do not achieve this level of perfection. Normally, a fixed number of bits (binary digits) is used to represent a sample value. Therefore, the infinite set of values possible in the analog signal is not available for the samples. In fact, if there are R bits in each sample, exactly 2R sample values are possible. For high-fidelity applications, such as archival copies of analog recordings, 24 bits per sample, or a so-called 24 bit resolution, should be used. The difference between the analog signal and the closest sample value is known as quantization error. Since it can be regarded as noise added to an otherwise perfect sample value, it is also often called quantization noise. 24-bit digital audio has negligible amounts of quantization noise.

Digital Image Capture

Digital image capture divides the image into a grid of tiny regions, each of which is represented by a digital value which records color information. The resolution of the image indicates how densely packed these regions are and is the most familiar measure of image quality. However, in addition to resolution you need to consider the bit-depth, the amount of information recorded for each region and hence the possible range of tonal values. Scanners record tonal values in digital images in one of three general ways: black and white, grayscale, and color. In black and white image capture, each pixel in the digital image is represented as either black or white (on or off). In 8-bit grayscale capture, where each sample is expressed using 8 bits of information (for 256 possible values) the tonal values in the original are recorded with a much larger palette that includes not only black and white, but also 254 intermediate shades of gray. In 24-bit color scanning, the tonal values in the original are reproduced from combinations of red, green, and blue (RGB) with palettes representing up to 16.7 million colors.

Digital Text Capture

Although it may seem odd to discuss digital text in this context, there are some important, if indirect parallels between the principles described above and those that govern digital text capture. Clearly in capturing digital text one does not sample the original in the same way that one samples audio or images. However, the process of text capture does involve choices about the level of granularity at which the digital representation will operate. In capturing a 20th-century printed text, for instance, a range of different “data densities” is possible: a simple transcription of the actual letters and spaces printed on the page; a higher-order transcription which also represents the nature of textual units such as paragraphs and headings; an even more dense transcription which also adds inferential information such as keywords or metrical data. Other possibilities arise in texts that have different kinds of internal granularity. In the case of a medieval manuscript, one might create a transcription that captures the graphemes—the individual characters—of the text but does not distinguish between different forms of the same letter (for instance, short and long s). Or one might capture these different letter forms, or even distinguish between swashed and unswashed characters. One might also choose to capture variations in spacing between letters, lines of text, and text components, or variations in letter size, or changes in handwriting, or any one of a number of possibly meaningful distinctions.

These distinctions, and the choice of whether or not to capture them, are the equivalent of sampling rates and bit-depth: they govern the amount of information which the digital file records about the analog source, and the resulting amount of nuance that is possible in reusing and processing the digital file.

NINCH guide table of contents

valid xhtml 1.1
abp~03/03

NINCH guide table of contents

References

American Memory Evaluation Team. “Final Report of the American Memory User Evaluation, 1991-1993.” 1993, American Memory Project, Library of Congress, http://memory.loc.gov/ammem/usereval.html, (acc. September 2002).

Bearman, D., G. Rust, S. Weibel, E. Miller, and J. Trant. “A Common Model to Support Interoperable Metadata. Progress Report on Reconciling Metadata Requirements from the Dublin Core and INDECS/DOI Communities.” *D-Lib Magazine* 5, no. 1, January (1999). http://www.dlib.org/dlib/january99/bearman/01bearman.html, (acc. Oct 2000).

Emery, P. "The Content Management Market: What You Really Need to Know." *Spectra*, vol. 29, no 1 (2002), pp.34-38.

Fraser, B. F. Bunting, and C. Murphy. *Real World Color Management*. Peachpit Press: forthcoming, autumn 2002.

Friedlander, A. “The National Digital Information Infrastructure Preservation Program: Expectations, Realities, Choices and Progress to Date,” *D-Lib Magazine* 8, no. 4, April (2002). http://www.dlib.org/dlib/april02/friedlander/04friedlander.html.

Gay, G., and R. Rieger. “Tools and Techniques in Evaluating Digital Imaging Projects”. *RLG Diginews* 3, no 3, June 15 (1999). http://www.rlg.org/preserv/diginews/diginews3-3.html#technical1 (acc. September 2002).

Hazen, D., J. Horrell, and J. Merrill-Oldham. *Selecting Research Collections for Digitization*. Council on Library and Information Resources (August 1998). http://www.clir.org/pubs/reports/hazen/pub74.html.

Johnson, N. F., and S. Jajodia. “Exploring Steganography: Seeing the Unseen.” *Computer* February 31, no. 2 (1998): 26-34.

Kiernan, K. “Digital Preservation, restoration, and the dissemination of medieval manuscripts.” *Gateways, gatekeepers, and roles in the information omniverse: proceedings of the third symposium: November 13-15, 1993*, edited by A Okerson and D Mogge, Washington, D.C., 1994.

Lesk, M. *Image Formats for Preservation and Access: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access*. Washington, D.C.: Commission on Preservation and Access, 1990.

Lesk, M. Report on *“Real World”* searching panel at SIGIR97. ACM SIGIR Forum, 32(1):1-4, Spring 1998.

Mintzer, F. C., L. E. Boyle, A. N. Cazes, B. S. Christian, S. C. Cox, F. P. Giordano, H. M. Gladney, J. C. Lee, M. L. Kelmanson, A. C. Lirani, K. A. Magerlein, A. M. B. Pavani, and F. Schiattarella. “Toward On-line, Worldwide Access to Vatican Library Materials.” *IBM Journal of Research and Development* 40, no. 2 (1996). http://www.research.ibm.com/journal/rd/mintz/mintzer.html, (acc. Oct 2000).

Mudawwar, M. F. “Multicode: A Truly Multilingual Approach to Text Encoding.” IEEE Computer 30(4): 37-43 (1997).

OCLC/RLG. “Preservation Metadata for Digital Objects: A Review of the State of the Art.” A white paper by the OCLC/RLG Working Group on Preservation Metadata, January 2001, http://www.oclc.org/research/pmwg/presmeta_wp.pdf.

Prescott, A. “Constructing Electronic Beowulf.” *Towards the Digital Library: The British Library’s ‘Initiatives for Access’ Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott. London: The British Library, 1998. 30-49.

Puglia, S. “The Costs of Digital Imaging Projects.” *RLG DigiNews* 3, no. 5, October 15 (1999).

http://www.rlg.org/preserv/diginews/diginews3-5.html#feature, (acc. Oct 2000).

Ross, S. “Strategies for Selecting Resources for Digitization: Source-Orientated, User-Driven, Asset-Aware Model (SOUDAAM).” *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 5-27.

Ross, S. and Economou, M. 1998. ‘Information and Communications Technology in the Cultural Sector: The Need for National Strategies’, *DLib Magazine* June 1998. http://www.dlib.org/dlib/june98/06ross.html.

Royan, B. “Cross-domain access to digitised cultural resources: the SCRAN project.” 64th IFLA General Conference, Amsterdam, 16-21 August 1998. http://ifla.inist.fr/IV/ifla64/039-109e.htm (acc. 8 March 2002).

Royan, B. “Scotland in Europe: SCRAN as a Maquette for the European Cultural Heritage Network.” *Cultivate Interactive* July 2000. http://www.cultivate-int.org/issue1/scran/ (acc. 2002)

Shapiro, M. and B. Miller. *A Museum Guide to Copyright and Trademark*. American Association of Museums, 1999. http://www.aam-us.org/resources/reference_library/mus_guide_copyright.cfm (excerpts).

Smith, A. *The Future of the Past: Preservation in American Research Libraries*. Washington, D.C.: Council on Library and Information Resources (CLIR), 1999. http://www.clir.org/pubs/reports/pub82/pub82text.html, (acc. Oct 2000).

Waibel, G. “Produce, Publish and Preserve: A Holistic Approach to Digital Assets Management.” Berkeley Art Museum Pacific Film Archive, n.d. http://www.bampfa.berkeley.edu/moac/imaging/index.html

Zorich, D. M. *Introduction to Managing Digital Assets: Options for Cultural and Educational Organizations*. Los Angeles: Getty Information Institute, 1999.

Zorich, D. “Why the Public Domain is Not Just a Mickey Mouse Issue,” NINCH Copyright Town Meeting, Chicago Historical Society, January 11, 2000. http://www.ninch.org/copyright/2000/chicagozorich.html

---

NINCH guide table of contents

valid xhtml 1.1
abp~03/03

NINCH guide table of contents

Abbreviations Used in the Guide

AAT	Art and Architecture Thesarus or Applications of Advanced Technologies Program
AHDS	Arts and Humanities Data Service
AIF	Audio Interchange File Format
AMICO	Art Museum Image Consortium
ASCII	American Standard Code for Information Interchange
BAMPFA	Berkeley Art Museum and Pacific Film Archive
BL	British Library
CCP	Classics Computer Project
CD	Compact Disc
CDL	Californian Digital Library
CDP	Colorado Digitization Project
CIDC	Cornell Institute for Digital Collections
CIMI	Consortium Museum Intelligence

CITI	Chicago Information, Tracking and Inquiry
CMS	Collection Management Services (Division, Library of Virginia)
CMYK	Cyan Magenta Yellow and Black
CNN	Cable News Network
CTIT	Centre for Telematics and Information Technology
DAT	Digital Audio Tape
DC	Dublin Core
DFG	Deutsche Forschungsgemeinschaft
DIMTI	Digital Imaging and Media Technology Initiative
DLIB	Digital Library
DLIT	Digital Library and Information Technologies
DLP	Digital Library Program
DLPS	Digital Library Production Service
DLXS	Digital Library eXtension Service
dpi	Dots per inch
DTD	Document-Type Definition

DVD	Digital Versatile Disc
EAD	Encoded Archival Description
EAF	Early American Fiction
FTE	Full Time Equivalent
FTP	File Transfer Protocol
G&M	Geography and Map (Division, Library of Congress)
GDZ	Göttinger DigitalisierungsZentrum
GIF	Graphics Interchange Format
GIS	Geographical Information Systems
GPS	Global Positioning System
GSU	Genealogical Society of Utah
HATII	Humanities Advanced Technology and Information Institute
HLF	Heritage Lottery Fund
HP	Hewlett-Packard
HTML	Hypertext Markup Language
IAIA	Integrated Arts Information Access project

IATH	Institute for Advanced Technology in the Humanities
IBM	International Business Machines
IMLS	Institute of Museum and Library Studies
IPR	Intellectual Property Rights
ITC	Information Technology and Communication
ITS	Information Technology Service
JPEG	Joint Photographic Experts Group
LAN	Local Area Network
LC	Library of Congress
LCSH	Library of Congress Subject Headings
LDI	Library Digital Initiative
LV	Library of Virginia
LZW	Lempel Ziv Welch
MAC	Apple MacIntosh Computer
MARC	Machine Readable Catalogue
MESL	Museum Educational Site Licensing

MIA	Minneapolis institute of Art
MIDI	Musical Instrument Digital Interface
MLS	Masters Library Science
MOA	Making of America
MOAC	Museums and the Online Archive of California
MOV/AVI	QuickTime Movie File Format
MPEG	Moving Picture Experts Group
MrSID	TIFF file viewer
MS	Microsoft
NDLP	National Digital Library Program
NEH	National Endowment for the Humanities
NMR	National Monuments Record (NMR
NRK	National Broadcasting Corporation (Norway)
NT	New Technology (Microsoft Operating System)
NTSC	National Television System Committee
NYPL	New York Public Library

OAC	Online Archive of California
OCR	Optical Character Recognition
OPAC	Online Public Access Catalogue
P&P	Prints and Photographs (Division, Library of Congress)
PAL	Phase Alternation Line
PC	Personal Computer
PDF	Portable Document Format
PEAK	Pricing Electronic Access to Knowledge
PM	Project Management
PT	Part Time
QC	Quality Control
QT	QuickTime
Ra	Real Audio/Video
RDF	Resource Description Framework
RGB	Red Green Blue
RSAMD	Royal Scottish Academy of Music and Drama

RTF	Rich Text Format
SCRAN	Scottish Cultural Resource Access Network
SGML	Standard Generalized Markup Language
SIBL	Science, Business & Industry Library (NYPL)
SLR	Single Lens Reflex
STG	Scholarly Technology Group
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format
TLG	Thesaurus Linguae Graecae
TML	Thesaurus Musicarum Latinarum
TOC	Table of Contents
UK	United Kingdom
ULAN	Union List of Artists Names
UMICH	University of Michigan
UNESCO	United Nations Education Social Commission
US	United States

USD	United States Dollars
USMARC	United States Machine Readable Catalog
UV	Ultra Violet
UVA	Virginia University
VRA	Visual Resources Association
VHF	Visual History Foundation
VIA	Visual Information Access
VSL	Virginia State Library
VTLS	Virtual Library System
W3C	The World Wide Web Consortium
WAC	Walker Art Center
WAN	Wide Area Network
WAV	Wave format
XML	Extensible Markup Language

valid xhtml 1.1
abp~04/02

NATIONAL INITIATIVE FOR A NETWORKED CULTURAL HERITAGE

ninch

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

Background Bibliography

Table of Contents

1. Project Management
2. Rights Management
3. Mark-Up (text; metadata)
4. Images (capture; management)
5. Audio (capture; management)
6. Moving Images (capture; management)
7. Preservation
8. Quality Assurance (content; metadata)
9. Asset Management
10. Collaboration
11. Case Studies
12. General
13. Standards

valid xhtml 1.0 strict

abp~10/02

NATIONAL INITIATIVE FOR A NETWORKED CULTURAL HERITAGE

ninch

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

By the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and NINCH

Interview Reports
Interviews conducted by The Humanities Advanced Technology and Information Institute (HATII)

1 Introduction

2 The Bancroft Library

3 Brown University, Women Writers Project

4 The Online Archive of California

5 The Art Institute of Chicago, Imaging Department

6 University of Chicago, The Oriental Institute

7 University of Chicago Library

8 Colorado Digitization Project

9 Library of Congress, Geography and Map Division

10 Library of Congress, Preservation Reformatting Division

11 Library of Congress, Prints and Photographs Division

12 Library of Congress / National Digital Library Program (NDLP)

13 Cornell University

14 University of Illinois, Digital Imaging and Media Technology Initiative (DIMTI)

15 Images of England project (National Monuments Record, English Heritage)

16 Göttingen State and University Library, Center for Retrospective Digitization

17 Harvard University, Visual Information Access (VIA)

18 Harvard University Fine Art Museums / Collection Management System

19 Harvard University Library / Library Digital Initiative Program (LDI)

20 Indiana University, Cook Music Library

21 The Church of Jesus Christ of Latter-Day Saints, Genealogical Society of Utah (GSU)

22 University of Michigan

23 Museums and the Online Archive of California

24 New York Public Library (NYPL)

25 National Library of Norway

26 Survivors of the SHOAH Visual History Foundation

27 Indiana University, Thesaurus Musicarum Latinarum (TML)

28 Tufts University

29 Library of Virginia (LVA)

30 University of Virginia, Center for Digital History (Alderman Memorial Library)

31 University of Virginia, Electronic Text Center

32 University of Virginia, Institute for Advanced Technology in the Humanities (IATH)

33 University of Virginia, Robertson Media Center (Clemons Library)

34 University of Virginia, Special Collections (Alderman Memorial Library)

35 University of Virginia, William Blake Archive

36 Walker Art Center (WAC) and Minneapolis Institute of Art (MIA)

37 WGBH Media Archives and Preservation Center & Digital Media Production

38 Appendix -- The Interview Instrument

valid xhtml 1.1
abp~09/02

The Interview Instrument

The interview instrument elicited spontaneous and favorable comment for its structure, comprehensiveness and clarity from a number of subjects. One interviewee commented that it was the best she had seen and another that it was a model. From an interviewer's perspective the instrument enhanced their ability to elicit good quality responses and sustain a logical chain of thought. Perhaps the best praise was that a majority of subjects commented that the interview had prompted them to think deeply and reflectively about their projects in ways that they had not previously done. Conducting the interviews face-to-face was certainly worthwhile. This was beneficial for the subjective impressions the interviewer gains using this method and the benefits of being able to get a "feel" for places and people. In addition, because the instrument needs to be standard to allow for comparable responses, yet cover digitization projects that use a variety of media and exist in a number of different institutional contexts, it inevitably has occasional shortcomings. Therefore, being able to offer an interpretation, explanation or example on a question where its meaning, relevance or application was not immediately apparent, was invaluable.

Nevertheless, one subject commented that they felt the instrument was more suited to libraries and that some areas could be addressed with more nuances for non-library projects. Certainly there are sections, particularly regarding cataloging and references to MARC, that are most applicable to libraries but it is difficult to see how this can be avoided if we are to gain information on their particular areas of expertise. Possibly an additional section with more detailed questions regarding teaching and learning or managing academic input would address this but it is straying off the purpose of the research. Furthermore, the size of the instrument is about as large as one could go, in most cases there was a sense of fatigue and loss of concentration towards the end.

The distinction at the start of the instrument over programs and projects, did not work in practice and was effectively ignored, leading to some repetition in questions. The latest version has been modified slightly to reflect this; another duplicate question was removed and the meaning of a couple more clarified (as much for the benefit of the interviewer as for the interviewee).

Here you can find a complete version of the Interview Instrument in pdf format. If you don't have the free Acrobat Reader to read it you can find this software here.

valid xhtml 1.1
abp~10/02

NATIONAL INITIATIVE FOR A NETWORKED CULTURAL HERITAGE
ninch

search

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

By the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and the National Initiative for a Networked Cultural Heritage (NINCH)

Acknowledgements

I. Introduction

II. Project Planning

III. Selecting Materials: An Iterative Process

IV. Rights Management

V. Digitization and Encoding of Text

VI. Capture and Management of Images

VII. Audio/Video Capture and Management

VIII. Quality Control and Assurance

IX. Working With Others

X. Distribution

XI. Sustainability: Models for Long-Term Funding

XII. Assessment of Projects by User Evaluation

XIII. Digital Asset Management

XIV. Preservation

Appendix A: Equipment

Appendix B: Metadata

Appendix C: Digital Data Capture: Sampling

References

Abbreviations Used in the Guide

Bibliography

Interview Reports and Interview Instrument

Click here to download a printable pdf version.

Please send us your comments and feedback!

For HATII

Seamus Ross
Ian Anderson
Celia Duffy
Maria Economou
Ann Gow
Peter McKinney
Rebecca Sharp

For NINCH

President: Samuel Sachs II
President-Elect: Charles Henry
Executive Director: David L. Green

NINCH Working Group on Best Practices

Chair: David L. Green
Kathe Albrecht
Morgan Cundiff
LeeEllen Friedland*
Peter Hirtle
Lorna Hughes
Katherine Jones
Mark Kornbluh
Joan Lippincott
Michael Neuman
Richard Rinehart
Thornton Staples
Jennifer Trant**

* through June 2001
** through May 1999

Copyright 2002, National Initiative for a Networked Cultural Heritage

Version 1.0 of the First Edition, first published October, 2002

I. Introduction

The Case for Good Practice

Early developers of digital resources often had little thought for how their projects might dovetail with others. Today many of these projects suffer from this lack of forethought; they cannot be extended for broader use, they cannot be built upon by others and the chances are slim that they will survive into the future. More recently, the cultural community has begun to realize the importance of applying technical and information standards intelligently and consistently. The use of such standards not only adds longevity and scalability to the project's life cycle, but also enables an ever widening public to discover and use its digital resources.

One of the goals of this *Guide to Good Practice* is to show the critical importance for the community of moving beyond the narrow vision of these early project-based enthusiasts and thinking through what is needed to establish sustainable programs. By adopting community shared good practice, project designers can ensure the broadest use of their materials, today and in the future, by audiences they may not even have imagined and by future applications that will dynamically recombine 'digital objects' into new resources. They can ensure the quality, consistency and reliability of a project's digital resources and make them compatible with resources from other projects and domains, building on the work of others. Such projects can be produced economically and can be maintained and managed into the future with maximum benefit for all. In short, good practice can be measured by any one project's ability to maximize a resource's intended usefulness while minimizing the cost of its subsequent management and use.

By adopting community shared good practice, project designers can ensure the broadest use of their materials, today and in the future, by audiences they may not even have imagined and by future applications that will dynamically recombine "digital objects" into new resources.

Within the cultural and educational communities, there are today many different types of guides to good practice written for particular disciplines, institution types or specific standards. These include the Text Encoding Initiative's Guidelines for Electronic Text Encoding and Interchange, Cornell University Library's *Digital Imaging for Libraries and Archives*, the Digital Library Federation's *Guides to Quality in Visual Resource Imaging*, the Getty Trust's *Introduction to Vocabularies* and *Introduction to Metadata* and the UK's Arts and Humanities Data Service series of discipline-based "Guides to Good Practice." In creating the National Digital Library, the Library of Congress has been assiduous in providing documentation and discussion of its practices; similarly, the National Archives has published its internal "Guidelines for Digitizing Archival Materials for Electronic Access," and the Colorado Digitization Project has brought together in a web portal a wide-ranging collection of administrative, technical, copyright and funding resources.

Link Box:

Existing Good Practice Guides

Guidelines for Electronic Text Encoding and Interchange (Text Encoding Initiative): http://www.tei-c.org

Digital Imaging for Libraries and Archives (Cornell University Library): http://www.library.cornell.edu/preservation/dila.html

Guides to Quality in Visual Resource Imaging (Digital Library Federation): http://www.rlg.org/visguides/

Introduction to Vocabularies (The Getty Trust): http://www.getty.edu/research/institute/vocabulary/introvocabs/

Introduction to Metadata (The Getty Trust): http://www.getty.edu/research/institute/standards/intrometadata/

"Guides to Good Practice" (Arts and Humanities Data Service): http://ads.ahds.ac.uk/project/goodguides/g2gp.html

"Guidelines for Digitizing Archival Materials for Electronic Access" (National Archives): http://www.nara.gov/nara/vision/eap/digguide.pdf

Various documentation from the Colorado Digitization Project: http://coloradodigital.coalliance.org/toolbox.html

The Library of Congress has published many supportive materials; some notable resources include:

"Challenges to Building an Effective Digital Library": http://memory. loc.gov/ammem/dli2/html/cbedl.html,

"Technical Notes by Type of Material": http://memory.loc.gov/ammem/dli2/html/document.html

"Background Papers and Technical Information": http://memory.loc.gov/ammem/ftpfile.html

"Manuscript Digitization Demonstration Project, Final Report": http://memory.loc.gov/ammem/pictel/

"Lessons Learned: National Digital Library Competition": http://lcweb2.loc.gov/ammem/award/lessons/lessons.html

"Conservation Implications of Digitization Projects": http://memory.loc.gov/ammem/techdocs/conservation.html

Put simply, this plethora of information is daunting. Where does one start and how does one evaluate the relevance of any particular text in the growing corpus of material on project planning, digitization, the kinds of metadata that need to be included in any project, and the maintenance and preservation of digital resources?

As we detail below, the *NINCH Guide* has a good claim to being unique in providing a broad platform for reviewing these many individual statements. First, it is a community-wide document, created and directed by a NINCH Working Group culled from practitioners from digitization programs in different types of institutions (museums, libraries, archives, the arts and academic departments) dealing in different disciplines and different media. Second, it is based on a set of broad guiding principles for the creation, capture and management of networked cultural resources. And finally, it is also based on a set of intensive interviews of substantial digitization programs in the U.S. and abroad. The perspective is thus a new one.

By offering universal access to the knowledge this research brings together, the *Guide* should help to level the playing field, enabling newcomers to the field and projects which are smaller, either in terms of budget or scope, to offer resources that are as valid, practical and forward-thinking as projects that are created within information- and resource-rich institutions. It is this sharing of knowledge that truly facilitates the survival and success of digital resources.

History, Principles and Methodology of the *NINCH Guide*

The National Initiative for a Networked Cultural Heritage (NINCH) is a US-based coalition of some 100 organizations and institutions from across the cultural sector: museums, libraries, archives, scholarly societies, arts groups, IT support units and others. It was founded in 1996 to ensure strong and informed leadership from the cultural community in the evolution of the digital environment. Our task and goal, as a leadership and advocacy organization, is to build a framework within which these different elements can effectively collaborate to build a networked cultural heritage.

Realizing from the start the importance of connecting the big picture (the overall vision and goals for a networked cultural heritage) with actual practice within cultural institutions, NINCH board and staff concluded that organizing a comprehensive Guide to Good Practice was an important priority. A NINCH Best Practices Working Group was created in October 1998 to organize a review and evaluation of current practice and to develop a set of principles and guidelines for good practice in the digital representation and management of cultural heritage.

The Group proposed an initial definition of good practice by distilling six core principles from their own experience with a set of evaluative criteria to judge current practice. The Group thus proposed that Good

Practice will:

1. ***Optimize interoperability of materials***

   Digitization projects should enable the optimal interoperability between source materials from different repositories or digitization projects

2. ***Enable broadest use***

   Projects should enable multiple and diverse uses of material by multiple and diverse audiences.

3. ***Address the need for the preservation of original materials***

   Projects should incorporate procedures to address the preservation of original materials.

4. ***Indicate strategy for life-cycle management of digital resources***

   Projects should plan for the life-cycle management of digital resources, including the initial assessment of resources, selection of materials and digital rights management; the technical questions of digitizing all formats; and the long-term issues of sustainability, user assessment, digital asset management and preservation.

5. ***Investigate and declare intellectual property rights and ownership***

   Ownership and rights issues need to be investigated before digitization commences and findings should be reported to users.

6. ***Articulate intent and declare methodology***

   All relevant methods, perspectives and assumptions used by project staff should be clarified and made explicit.

With funding from the Getty Grant Program, NINCH issued a request for proposals to conduct a survey and write the *Guide*, in close collaboration with the Working Group. A team organized by the Humanities Advanced Technology and Information Institute (HATII) of The University of Glasgow was hired.

In order to ground the *Guide* in the reality of good practice that has been proven in the field, and to ensure that the personal views of the Working Group did not color the guide too much, the project began with a thorough review of current literature on the subject of good practice that included online and print resources, as well as gray[1] literature. This process was complemented by structured face-to-face and telephone interviews, and selective written exchanges with individuals from the cultural heritage sector.

The key information-gathering tool used for research was the *Digitization Data Collection Instrument for Site Visit Interviews* developed by HATII. For details on the development and use of this interview instrument see the "Introduction" to the Interview Reports. Interviews at digitization facilities lasted between 90 minutes and 3 hours and were conducted by four researchers on 20 site visits, involving 36 projects and 68 individuals from late 2000 through early 2001.

Sites were selected on a "best fit" basis to a matrix of project types and key themes established by the project team. The sites selected were not a scientific or representative sample, but as a group they broadly reflected the diversity of the community, while each represented one or more of the identified key themes of good practice. The rationale for site selection is further explained in the "Introduction" to the Interview Reports.

In parallel to the site visits, the research team undertook further focused research via literature review, telephone interviews and written correspondence on several broad themes: text encoding, digital preservation, asset management, rights management, and quality assurance. HATII identified another set of relevant digitization sites for inclusion in this stage of research. Theme reports written out of this research filled knowledge gaps that had not been addressed by the site visits and provided a more analytical view of current community good practice in these areas.

How To Use the *Guide*

The *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* is a unique contribution to the field. It takes a process-oriented approach to the digitization and management of cultural resources (keeping in mind their long-term life cycle from selection through preservation) and does so from a community-wide perspective. NINCH also intends to put into place a system for regular updates and further editions. The *Guide* takes the reader from the identification of available resources and the selection of material, through the creation of digital content, to its preservation and sustained access. For institutions that have not yet begun digitally representing material from their collections or making their born digital material accessible, the *Guide* will provide a way of coming up to speed in a quickly developing area. It identifies the decisions that need to be made, indicates when they need to be made and draws attention to the implications of the possible choices.

Users of the *Guide* will come from different backgrounds. Perhaps five examples will help you situate yourself among the possible categories of readers.

- If you are an archivist, librarian or museum professional, the *Guide* will help you select materials from your collections, reformat them, and make them visible and accessible to different audiences via the Internet or on portable digital media.
- If you are a funder, the *Guide* will give you an understanding of the activities involved in creating, delivering and sustaining digital content and background, and will help you to assess whether or not requests for funding are sensible and built on a thorough consideration of the issues.
- If you are an academic or other researcher, the *Guide* should give you sufficient information to design a project, convince collection owners to grant you access to material you need to digitize, and persuade funders to support your project.
- If you are a teacher of digitization in a library school or a faculty of information studies, the *Guide* can help you identify central issues to cover in digitization courses, and can provide your students with an understanding of the issues that they will need to address when they join a cultural heritage institution.
- If you are a vendor or manufacturer of software or hardware, the *Guide* should provide you with an indication of the challenges faced by the cultural community and of the significant levels of investment that the community is making in digital content creation, as well as showing you the tremendous value of the intellectual capital with which they are working.

This is not a recipe book for experts or specialists. It will provide content owners and decision-makers with sufficient guidance to know whether or not they are getting the best advice from their technical staff and whether their colleagues have put in place adequate strategies to guarantee the success of their digitization activities. It does not attempt to provide the final word on every topic, but instead supplies links to resources that we have evaluated and have concluded will offer a good next step.

The *Guide* identifies the decisions that need to be made, indicates when they need to be made and draws attention to the implications of choices made.

Humanities and cultural heritage institutions serve the needs of many different communities - from students and scholars to publishers and the general public. As you begin to develop and plan the use of digitization to make your collections visible and accessible, it is crucial to decide which audiences you aim to reach. This will influence many of your decisions: the items you select for digitization, the technologies you will use, and the mechanisms for delivering the digital materials to users. You may find, for example, that you have a collection that interests children as well as adults, but that each audience will require different delivery interfaces. While you could use the same technologies to reformat the material (and you would only need to do it once), and publish both versions using the same underlying delivery system, you would have to develop two separate interfaces to the same material.

Digitization may even change your sense of audience, by making it possible to offer broader access to rare or

inaccessible collections. Institutions often think first about digitizing material that is already popular with the public, but digital technologies now enable them to offer access to material that could not otherwise be seen or used, thus altering rather than simply reproducing the existing profile of use.

Audiences may be not only the users of the digital collections you produce, but also potential creators of digital surrogates from your collection for research, publication, advertising or enjoyment. Examples might be:

- an academic asking to digitize a collection of papers by a recently deceased contemporary artist as part of a research project
- a publisher proposing to produce a pay-per-view website with images of your collection of sixteenth-century engravings of native Americans
- a folk society requesting permission to include a rare recording of a 20th century storytelling from your collection on a CD they hope to release.

How do you respond to these requests?

- What best practices would you require if you were to agree to any or all of them?
- Would your expectations of each project be different or would you set them the same high standards?
- How would you ensure that, while you allow them each to use the material for their different purposes, you retain control of it in digital form, and that the processes involved in its digitization do not put the analog material at risk?

It is worth remembering that analog holdings constitute intellectual capital, and that as digital surrogates are created, the research, teaching or economic value of the originals should not be depleted. This may affect the material you choose to make accessible, the standard to which you do so, and what types of use and access arrangements you will put in place. Requiring those who work with your collections to follow good practices can minimize risks to the analog sources through their digitization.

So the first questions to ask include:

- Where is the audience for my collections?
- What types of individuals does that audience include?
- Will digitization enable me to meet the needs of existing communities better?
- Will digitization enable me to create new audiences for both the digital surrogates and the analog sources?
- What do I mean by “audience” in the digital world? Am I referring only to those individuals to whom I can deliver digital materials or am I also giving consideration to those who would like to produce digital surrogates for business, personal and research purposes?

---

[1] Gray literature, sometimes called "ephemeral literature," is unpublished material that can be lost to potential readers because it is not disseminated widely through publication or indexing. Examples of gray literature include: government or NGO research reports, workshop or conference papers, and theses.

Guide to Good Practice in the Digital Representation of Cultural Heritage Materials - Interview Reports

Introduction

Data Collection Instrument

In tackling the research for the Guide, the research team hired by NINCH, from Glasgow University's Humanities Advanced Technology and Information Institute (HATII), established an empirical methodology that adopted four main instruments: literature reviews, face-to-face interviews and observation, telephone interviews and written discussion.

The key information-gathering tool used for research was the *Digitization Data Collection Instrument for Site Visit Interviews*. In developing this instrument, the HATII research team drew on the six principles developed by the NINCH Working Group in addition to building upon a number of previous surveys.[1] Consisting of six main sections, made up of 31 parts and 202 questions, the instrument was designed to collect data on all types of media (still and moving images, text and sound) and from all project types (projects or programs, large, medium or small scale, museum, gallery, library or academic based, publicly or privately funded, etc.). The site visit instrument was used in preference to free-form interviews in order to provide researchers with a core of comparable data irrespective of the interviewer, time or place of the interview or project type. It was used by four researchers on 20 site visits, involving 36 projects and 68 individuals from Fall 2000 through Spring 2001. The instrument proved able to elicit high quality responses and sustain a logical chain of thought. In many cases it prompted subjects to think deeply and reflectively about their projects in ways that they had not hitherto done.

The Interviews

The HATII team considered these extensive face-to-face practitioner interviews as an essential part of the survey of the field. It placed people and projects in context, enabling interviewers to observe facilities, equipment and practices first hand. Standardized to allow for comparable responses, yet covering a variety of media and projects, the interview elicited many responses that inevitably required interpretation or explanation where their meaning, relevance or application was not immediately apparent. Interviews lasted between one hour 30 minutes and three hours and it was found that the instrument was as long as can be used in one session.

Subjects were sent a copy of the instrument in advance to lessen the chance of their being unable to answer questions, (thus improving the quality of data) and to build a sense of openness and trust — promoting the fact that the interview was an information gathering exercise and not a critical assessment.

Each of the interviewers used the instrument to structure the interview and gather as complete a profile of each project as possible. An annotated transcript of each interview was subject to internal review before being sent to each subject for comments, corrections or additions. At this stage, the transcript was intended to represent as faithfully as possible the interview itself and provided an opportunity to ask any follow-up questions or request clarifications. Upon receipt of any changes by the interviewer, the transcript was re-written and subject to a second internal review. The transcripts were then amalgamated into a single document and submitted to members of the NINCH Working Group. Following their review and amendment, a draft publication copy of the transcript was sent to subjects for final follow-up questions and review. The transcript was then amended and given its third internal review before being incorporated into the draft Guide.

Other Research

In parallel to the site visit data collection, the research team also undertook focused research on a thematic basis with the goal of expanding data collection in five particular areas: text encoding, digital preservation, asset management, rights management, and quality assurance. This research was conducted by a combination of literature review, telephone interviews and written correspondence, supplemented by specific questions for key practitioners in the relevant field. In addition to the sites selected for site visits, the project team identified a further series of relevant projects for inclusion in this stage of research. This research produced a series of theme reports that filled knowledge gaps that had not been addressed by the site visits and provided a more analytical view of current community good practice in these areas.

Site Selection

In addition to data collection methods, site selection was the second major methodological issue faced by the project. Sites were selected on a “best fit” basis to a matrix of project types and key themes established by the project team. The digitization and networking of cultural heritage materials takes place across a variety of media and in a variety of institutional contexts. In order to reflect this diversity, the research team extensively reviewed potential sites. Collectively, the sites needed to incorporate projects that digitized still images, texts, moving images and audio as well as education, library, museum, gallery and archive institutions. However, the project did not select sites to reflect the proportion of each media type being digitized or the extent to which particular institutional types are represented in the cultural heritage community, even if such figures could be ascertained. In this sense the sites selected were not a scientific or representative sample, but ones that broadly reflected the diversity of the community, while representing one or more key themes of good practice.

Since the aim of the Guide was to seek out and document good practice rather than merely profiling current practice, the project team's expertise played a crucial role in identifying sites that would support this goal. Particular care was given to establishing an open and transparent selection process, a stringent review of proposals, and a project team with broadly international expertise in cultural heritage digitization projects.

---

[1] - The Federal Task Force on Digitization (Canada). *Towards a Learning Nation: The Digital Contribution.* http://www.nlc-bnc.ca/8/3/ report.txt
- Internation Federation of Library Associations and Institutions. *Worldwide Survey of Digitised Collections in Major Cultural Institutions: an IFLA PAC/UAP Joint Project* http://www.ifla.org/VI/2/ p1/miscel.htm
- InterPARES Project. *Authenticity Task Force: Case Study Protocol Questionnaire* http:// www.interpares.org/documents/interpares_CSIP.pdf
- S.D. Lee, *Scoping the Future of Oxford's Digital Collections.* Bodleian Library, Oxford University. http://www.bodley. ox.ac.uk/scoping/report.html
- University of Glasgow: Humanities Advanced Technology and Information Institute. *Funding Information and Communications Technology in the Heritage Sector Policy Recommendations to the Heritage Lottery Fund* http://www.hatii. arts.gla.ac.uk/Projects/HLF/

valid xhtml 1.1
abp~10/02

22 University of Michigan

On October 16 2000, HATII interviewed John Price Wilkin, Head of the Digital Library Production Service at the University of Michigan; Christie Stephenson, Assistant Head; and Maria Bonn, Head of the Scholarly Publishing Office. The Digital Library Production Service was initiated in 1996 with the intention of creating an infrastructure for the digital library collections at the University. One of a number of units within the University Library's Digital Library Initiatives division, the DLPS is responsible for the operation and maintenance of current collections, while also preserving these resources for future use.

22.1 Organizational Digitization Program and Policy

The Digital Library Production Service (DLPS) has not carried out systematic collection surveys to identify materials for digitization. Units within the library with collection responsibility drive the digitization process. Systematic surveys may be done for some projects which are targeting a segment of a collection. Cataloging is performed by staff in the Cataloging department. Preservation staff participate in projects like the Making of America (MOA) by ensuring that the materials are not "last copies", that they are processed as preservation materials and that they are appropriate for digital capture. Preservation staff conducted a collection survey before including materials in MOA 4. The overall focus is on the project, which can have different motivations, for example to clear shelf space or to digitize a particular subject.

The priorities are driven by the material — e.g. shelving, cataloging, condition, demand — and collection management is undertaken by collection development staff. The priorities are not formalized, and they can change through the digital process; for example, MOA revived interest in materials that had not attracted attention for many years. DLPS plans to develop selection criteria for conversion projects; this will not be prescriptive but will take the form of a framework.

The program recommends standardizing workflows for the digitization process and for quality assurance. They also recommend creating strategies with collection managers to ensure the matching of materials, methods and goals.

The obstacles identified by DLPS in development are mainly in connection with the data and methodology. In some projects it has been difficult to obtain the data. Some projects have difficulty with the methodology in working with the data, e.g. working with newspapers, and the judgements required in order to create usable resources.

The complexity of the workflow was identified as a main obstacle in the process. DLPS would ideally like to bring systematic thinking to the process to adapt to problems of scale. For example, MOA had to overcome issues associated with co-ordinating 27 staff spread over the library to achieve the goals of the project. The workflow also had to deal with vendors, outsourcing and complex project management systems.

There is a tension between costs and methods, and scaling projects, such as how to put small-scale projects into large-scale workflow.

The main criteria for the selection of material for DLPS are research significance and enhanced access. Other drivers are: intellectual property rights; teaching and learning potential; conservation; historic and cultural value; improved functionality; space rationalization; and preservation. By-products of the process (but not full criteria) are publicity, potential commercial exploitation and support of infrastructure costs. Other factors include provision of user services, labor cost reduction and social inclusion.

The conversion strategy varies from project to project and has changed over time. MOA 4 has different drivers from the first MOA project — moving from hand selection to automated selection.

In developing the digitization program co-operation has taken place with libraries, museums, corporations, foundations, government agencies, and academic institutions. The co-operation was local, regional, national and

international. DLPS feels that co-operation is a valuable experience but must be thought out carefully beforehand. It would encourage any institution to speak to other similar organizations before starting any digital program and to set communications lines and shared goals, giving one person the authority to enforce milestones and timelines. There should be a process of accountability, but appropriate level of communication.

The current status of the program is ongoing with no anticipated end date. Digital projects have been active in the University of Michigan (UMICH) since the late 1980s, digital library initiatives since 1993, Humanities Text Initiative (HTI) since 1994 and MOA since 1995. DLPS was established in 1996.

The main purposes in creating the digital deliverables for DLPS are preservation, public access, research, and wider access. Other purposes are teaching and learning, and revenue generation.

Specific projects have created statements of intent, and the content will vary from project to project.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- typescript documents;
- unbound printed documents;
- hand-written documents;
- printed illustrations;
- intermediaries;
- mechanical copies;
- photographic prints;
- photographic negatives;
- archaeological artifacts;
- pottery shards;
- ostraca;
- maps;
- specimens;
- papyri.

The nature and format of the materials digitized vary considerably from project to project and include all sizes of photographs, original artworks and various sizes of documents. Depending on the specific project, they may represent the whole collection or a sample.

DLSP intends its digital objects to be used again in a variety of projects including printing and for others to use.

The following standards, guidelines or tools are used for representing content:

- TEI;
- Unicode (one off and working on);
- SGML;
- XML;
- TIFF;
- JPG.

The following standards, guidelines or tools are used for describing content:

- MARC;
- Dublin Core (mapping strategies);
- CIMI;
- VRA Core;
- CDWA (perhaps);
- EAD;
- TEI header.

The following standards, guidelines or tools are used for controlling data values:

- LC subject headings;
- LC thesaurus for Graphic Materials;
- AAT;
- Name authority files.

The following standards, guidelines or tools are used for representing structure:

- TEI;
- XML;
- SGML;
- Numerical Data Rectangle (a statistical package).

DLPS looks at appropriate standards and guidelines for each project. Many of the standards available were not necessarily suitable for its needs, or in some cases the time and expertise were not available. Anne Kenney's guidelines on page imaging were extremely valuable (Kenney, Anne R. *Digital Imaging for Libraries and Archives.* Cornell University Library, June 1996).

Many guidelines were adapted for projects, such as the papyri project.

Use of controlled vocabularies and striking a healthy balance between standards and reality are both recommended: standards should not be used for their own sake, and pragmatic solutions are often the most effective.

The intended audience for digital deliverables, with their priorities, are:

- K-12 (low priority);
- community college (low priority);
- four-year college (high priority);
- graduate school (high priority);
- lifelong learning (low priority);
- distance learning (low/medium priority);
- museum users (medium priority);
- public library users (medium priority);
- archive users (medium priority);
- papyrologists (medium priority).

DLPS identified the academic sector as its target audience. It conducted evaluations on some projects' target audiences, for example the papyri project and MOA.

Other users whose profile will vary depending on the material can use all the objects.

W3C's "Guidelines for Web Site Accessibility" were considered for use in the systems. The project stated on the web page reasons for limitations of use, which were mainly due to costs and the nature of the material.

It is difficult to gauge if the target audience differs from the actual audience. Feedback is provided via web logs and via email.

22.2 Project Management and Planning

The DLPS uses in-house expertise for project management. The project management is tied to the cataloging and preservation department within the larger structure of the library. The work around digital library issues is changing the structure of the organization. For example, the preservation department has changed some of its staffing as a result of the collaborative work with DLPS on projects such as MOA.

Formal project management procedures are in place. The DLPS administration group provides online guidelines and procedures, and oversees project management. Procedures have had to be adapted to strike a balance between overly centralized procedures and insufficient project management. They have tried to be inclusive of individuals.

There is no formal quality assurance for project management.

No pilot study or feasibility study is carried out for the projects. There is a form of "ramp up" in projects when test scans are carried out. This is an evolving practice and they continue to learn.

Microsoft Project and flowcharts have been used to aid planning. Work allocation depends on previous responsibility. DLPS has produced procedural documents, which cover OCR bitonal and continuous tone scanning, quality assurance and preparation.

The digitization is carried out both in-house and externally. The method used depends on the materials, with digitisation of some fragile material being kept in-house. Cost and scale are factors in this decision. OCR tends to be done in-house, with encoding and bi-tonal scanning typically (but not always) outsourced. In-house methods enable further training of staff and control over the material and process.

The institution bought equipment for the digitization process. The capture process is driven by the material and project goals. Flatbed scanners, film scanners and high-end digital cameras are used. A transparency film scanner (Imacon Flextight Precision II Film Scanner) scans 4x5 transparencies; it bends the transparencies so that, although not a drum scanner, it is more than a slide scanner.

They are constantly working on the guidelines for data capture procedures and use gray scales, color bars and reproduction charts.

The image capture process consists of flatbed scanning, slide scanning, direct digital capture (Kontron) as well as photographing the object, producing 4x5 slides transparencies and digitizing these. The resultant digital objects are considered to be superior to the ones produced by the high-end digital cameras.

22.3 Human Resources and Training

The organization is divided into three functional groups: Digitization, Information Retrieval and Architecture. There is currently one director, an assistant director, five FTE digitization staff (supplemented by students), two collection co-ordinators for information retrieval, six programmers and a programming manager, a data loader/media manager, an interface specialist, a technical support person, and an office support person. The staff have a variety of backgrounds and general education. Staff are employed to work in DLPS and are not redeployed from other jobs within the Library. They use both external and in-house advice on the technical aspects of digitization. Specialist technical and equipment operators in bitonal scanning, OCR, text encoding and continuous tone image capture are trained in the following areas:

- project management;
- technical standards;
- preparation;
- operation;
- post-processing.

The training is provided in-house, through external courses, contact with peers in other institutions and by learning on the job.

22.4 Project Life-Cycle Processes and Procedures

DLPS is aware of the copyright status of its materials, but the institutions for which it manages content are ultimately responsible. The copyright declaration is in a text statement and sometimes on the image. The project digitizes copyrighted material under library provision and with the owner’s agreement. The users can make local copies of content for subsequent reuse (print, copy into another program, etc.). Downloading to a PC, LAN or WAN is permitted. Users can download and view ASCII text, encoded text or PDF. For images, users can download and view thumbnails, lower quality images, highest quality images and associated descriptive metadata and documents.

Watermarking is not used as this is considered to be ineffective and costly and as yet there is no system that actually works well. There is no recognized standard format and only proprietary systems are currently available.

UMICH has a preservation/conservation procedure for original material. The library preservation department conducted the brittleness survey for MOA; they also examine materials at high risk. Conservation may be carried out as part of the overall workflow, prior to actual digitization.

The risks identified to the material are handling, heat, breakage and dis-binding. Workflows and methods have been adjusted to lessen the risks and special equipment such as cold lights, cradles and exposure limits is used. Where necessary, curatorial staff prepare the material and also monitor the materials at the start of the process until they are happy with procedures. Availability of digital files does not preclude access to original materials, but it will be discouraged, and users encouraged to use the digital objects.

DLPS has always had access to all the relevant material although it sometimes has had to locate some additional material for projects. It has rejected material that has been too brittle. DLPS digitizes both originals and intermediaries.

For text conversion projects, DLPS has digitized original printed material, microfilm and occasionally photocopies (for replacement pages). For image conversion, DLPS captures originals including photographic prints, negatives, 35 mm slidesand creates 4x5 film intermediaries for some capture.

UMICH libraries use the full MARC record catalog and descriptive information is always taken from this to create headers for titles being digitized. UMICH libraries' original objects are cataloged by the appropriate curatorial unit (e.g. special collections or by the cataloging department). A document is forthcoming that explains the policy for cataloging digital surrogates.

MARC, EAD and USMARC are used to describe text and finding aids and to create collection level records for image collection. Images cataloged within the library employ a locally developed system called VRO (Visual Resources Online). Tools used for controlling data values are LC subject headings and catalogers desk top, as well as picklists and TGM values. The system they use has controls to guard against user error in input.

The projects record metadata about the original object, the digital object, the digitization process, technical details and administrative data. They are currently working on recording staffing details. They also have information inserted in the TIFF header by the vendors. The digitizer and the information professional create the metadata. The final metadata are included in the main catalog, except for visual resources that are kept in a separate catalog. The main catalog was in paper form until 1989 and is now available on the local intranet and the internet. URL and USMARC 856 link the digital and original objects.

22.5 Format, Resolution and Compression of Digitized Materials

DLPS uses SGML and TEI, XML, GIF and PDF for Page Image delivery, and TXT for delivery from SGML. Many of the documents have non-Latin characters, e.g. Greek. They have used beta-code with character entity reference in order to render them for display. They have used various OCR packages over the years but now use PrimeRecognition software to convert page images to text. For OCR without correction, they achieve approximately 99.8% accuracy or better. For OCR with correction, the accuracy is 99.995% (i.e., 1 error in 20,000), determined by sampling, otherwise the text is subjected to another round of proofreading. For keyboarding, they pay for (and verify that they have received) 99.995% accuracy. The documents do not have any special treatment prior to OCR. The aims are to create automatic indexing, enhanced searching, computer based analysis and document retrieval.

Experience has shown that converting a large volume of texts ensures a drop in costs. They recommend that co-operation is very useful to share facilitates as well as to increase the volume being processed. They also recommend buying and testing OCR packages to find the one most suitable for the type of text to be digitized.

DLPS also used keying in as a method to retroconvert text, which gives greater accuracy, but the specifications must be clear. Companies have responded favorably to this and this method is used where it would be absurd to try to OCR the documents.

TIFF file format is used for capturing and preserving scanned images. GIF and JPEG are used for delivery, along with PDF, Wavelet Compression and MrSID. They are seriously considering PNG.

Capture resolutions vary according to source and method. In general, page images are captured at 600 ppi and saved as G4 bi-tonal TIFFs. Resolutions for grayscale and continuous tone images vary depending on the size and nature of the original object and the capabilities of the capture device, but are saved as 8 and 24-bit TIFFs respectively.

For continuous tone images, DLPS always retains the original scans in uncompressed form as TIFFs. Images are stored as Mr. SID files (wavelet compression) and JPEGs are delivered to end-users. JPG is used to compress the files, aiming to improve access, enhance usability and decrease storage requirements.

Color correction, cropping and de-speckling are carried out using a variety of tools post-process. ImageMagic is used to create thumbnails. The file sizes vary hugely, from approximately 120K to 124 MB.

DLPS feels strongly that digital cameras are not the only viable alternative for digital image capture, and recommends that projects are careful to start with the materials and not the equipment. They have had great success with film intermediary as a method of image capture, with subsequent digitization done with a high-end transparency scanner. They argue for the better use of vendors and the establishment of community based, non-prescriptive open dialogue. Investment must be made to ensure success.

DLPS has not yet digitized a significant amount of sound or moving image material, but has begun serious exploration of support for both.

Quality control spot checks happen naturally through the process. They also have checks on random samples. Total checks are done in-house on continuous tone objects. On the papyrus project, the papyrologists organized the quality control. Outsourced material is quality controlled by the vendors as well as internally. For bi-tonal scanning, 5% of the images are checked for legibility and acceptable skew. ImageTag, the program used to assign page-level metadata values to the image files, provides a more thorough check, as an operator scans the images to link sequence to pagination information. The more formal quality control has helped the process; former projects had no such quality control and suffered accordingly.

Access to the digital objects varies with the copyright or license-status of the material or the preference of the image owners. In some cases, external users have only limited access to lower quality images. In-house users have access to the entire catalog and objects. Users can run a full text and fielded search on the associated image metadata and browse the image data. Users do not need any other software or hardware in order to access the materials. They can create image class sets, zoom and pan the images. The usage varies (see http://stats.umdl.umich.edu for more accurate information). The usage data is captured using formal methods and informal observation.

Users have to pay for the use of some of the digital deliverables. Some of the materials have a license to which users have to subscribe for access; this also includes UMICH. The publishers set the costs, and there is different pricing for the size of institution. If the subscriptions reach an optimum level then a price drop is considered (although this scenario has not actually happened).

Costs are not a driver in the projects; some are calculated on the usage and some are projected.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, print and broadcast media coverage, conferences, meetings and electronic and conventional mail shots. The project believes that listservs have been very successful in disseminating the information on MOA, while the academic articles are useful in communicating the overall programmatic methods and goals.

22.6 Evaluation, Funding and Long-term Sustainability

DLPS has run front-end evaluations using formal methods, online questionnaires, email, focus groups, and user observation. These evaluations have provided input into the development cycle, producing an iterative process.

Formative evaluation takes the form of questionnaires online, email, focus groups and recording interaction logs as well as observations of users' interaction, which is videoed for later reference. The resulting changes have been part of the iterative process.

One summative evaluation has been carried out, as all else is ongoing. The project was Pricing Electronic Access to Knowledge (PEAK) and included the summative evaluation component as part of the project when it was established. The project used questionnaires online, focus groups and computer interaction logging as evaluation methods.

DLPS has decided to make the evaluation process more formal and invite users to come in for one day a year to look at a variety of projects and areas. It is hoped that this will help steer the development in a more formal manner. It is difficult to get people to help with evaluations, and so they will offer incentives for users to attend this one-day event. The results from this event will be made available to personnel involved in information retrieval design.

The funding for DLPS comes from general funds from UMICH with a small amount of government and federal contribution. They feel adherence to standards has saved money in the long-term. Any documentation required will vary from project to project. Most documentation is part of the general process.

DLPS will continue to add new materials, update metadata, add metadata and amend the user interface. This is an ongoing process. User interface changes are made periodically, with some review done annually. File formats are changed only when the technology dictates. They have formal plans to document the digitization process for long-term preservation. They emphasize that any strategy is based on the materials and not the systems, so migration of data is important, but retention of hardware/software is not.

They intend to keep the digital deliverables available for as long as possible and the loss of this material would matter greatly.

The material, and not the equipment, should lead the process; where possible, scaleable extensible systems should be built to allow flexibility. This is happening through the Digital Library eXtension Service (DLXS - www.dlxs.org) - DLXS is the underlying system for delivery of DLPS digital library materials and is also offered to other non-profits as a supported combination of search and retrieval software and middleware for delivering digital library information including bibliographic, text, and image data.: “The University of Michigan Digital Library eXtension Service (DLXS) provides the foundation and the framework for educational and non-profit institutions to fully develop their digital library collections.” (see http://www.dlxs.org/aboutdlxs.html for a fuller overview).

valid xhtml 1.1
abp~04/02

15 Images of England project (National Monuments Record, English Heritage)

Vikki Fenner, the Project Manager for Images of England, and Phil Turton, the Technical Manager, were interviewed by HATII on February 16 2001. As part of the National Monuments Record, the Images of England project aims to encourage knowledge and enjoyment of English Heritage sites. The project was therefore undertaken in order to provide public access to the archives as well as documenting and preserving sites of historical and cultural value. Through this project, the NMR is building a digital library of photographs of 370 000 listed buildings in England.

15.1 Organizational Digitization Program and Policy

Images of England is managed by the National Monuments Record (NMR), the public archive of English Heritage. The project has mobilized volunteer photographers, the majority of whom are members of the Royal Photographic Society, to capture an image of every listed building in England. An initial pilot was carried out in 1996. They received the first phase grant from the Heritage Lottery Fund (HLF) in August 1999 and have a completion date for October 2002. The practice of listing buildings to preserve and protect them for future generations began in 1947 in Britain when the importance of preservation was realized after years of wartime destruction. Listed sites include palaces, castles, churches, country houses, as well as garden sheds, fishponds, pigeon lofts, mile posts, army camps, telephone boxes, pigsties, post boxes, and lamp posts. A prototype is now available on the website and the whole database will be available on the Internet by 2002. The project will generate around 370,000 new images. The aim is to create a ‘point in time’ record of English heritage, which will include buildings from every town and village.

15.1.1 Collection Survey and Priorities for Digitizing

They did not carry out a collection survey, as they already had a complete database with information about all the listed buildings in England from the NMR and so knew what they had to digitize. Setting priorities for digitizing also works quite differently in their case, as the staff are all volunteers. It is therefore the responsibility of the volunteer photographers to select the ‘defining’ view of the listed buildings in their area, following the training and guidelines set by the project team. Due to the voluntary nature of their work, the project team no longer gives them targets on how quickly they should process material. They tried using strict deadlines in the first year, but this proved unsuccessful, so they changed the approach in response to feedback from the volunteers themselves. There are about 400 volunteers who receive a list of ‘target’ buildings in their area, and the time commitment to the project and schedule planning is up to them, although they are generally encouraged to commit at least two days a month to the project. They used the pilot project loosely for planning purposes, to estimate for example the rate at which material was being photographed.

15.1.2 Overall Obstacles

One of the obstacles to both planning the development and building the digital deliverables, was that they had not anticipated that they would need a database in place to manage and monitor the survey, so they had to design it themselves after the start of the project. There was a time lag between putting in the HLF application and receiving the money. In addition, the grant was received in two stages, with a delay between them. The organisations did not have experience working with so many widely distributed volunteers in such a concentrated way and faced initital challenges because of this lack of experience. In practice it also proved quite resource-intensive, as they had more volunteers than they originally anticipated. Technology was not a problem in their case. In fact, they benefited from starting a little later, when technology was established in this field.

15.1.3 Selection and Prioritization of Materials

The project will create a ‘defining image’ of the exterior of each listed building in England. This needs to capture the essence of the architectural character of the building and something of its function and context. Photography is only taking place from publicly accessible land, unless the owners have given permission to photograph from privately owned land. The remaining selection and prioritization questions in the instrument do not apply in their

case.

15.1.4 Co-operation

There was wide-ranging regional and national co-operation with other organizations including archives (internally, as part of English Heritage), libraries, academic institutions, and corporations (as part of the consultation process and at the road shows organized across the country), foundations and charities, and government agencies. This experience showed that in any collaborative project involving volunteers, the organization should consider the population to draw upon very carefully.

15.1.5 Purpose of the Program

The aim is to create a 'point in time' record of English heritage, which will include buildings from every town and village. Between 1999 and 2002 volunteer photographers, the majority of whom are members of the Royal Photographic Society (RPS), will capture a single defining image of the exterior of every listed building in England. These 370,000 images will then be linked to existing listing information about the buildings, creating a digital library of England's architectural heritage.

In order of priority, the project manager would say that the project's purpose focuses on the following areas:

- preservation and public access;
- teaching and learning resource;
- research;
- experiment;
- revenue generation.

A statement of intent was produced with the application to the HLF. They also state the aim of the project on the website. This was explicit about the project's rationale, the scope, the significance, the primary audience, the long-term sustainability, and the suitability for different target audiences.

15.1.6 Nature of the Source Materials and the Impact of Digital Deliverables

The program digitizes only 35 mm color photographic negatives of all the listed buildings (so the entire body of the material in that sense). The digital deliverables are intended to be re-used. After the whole library is created, there will be spin-offs; for example, they might produce educational packs, both hard copy and digital, as well as sell copies of the images commercially.

15.1.7 Interoperability

The digitized images will complement the existing database of listed buildings. This uses various standards for cataloging: e.g., BS76666, NGDF (spatial data standard), and ANSI Z39.50. The information that the project is adding includes fields such as the name of the photographer, the date photograph was taken and the direction the camera was facing. They have standardized the way these should be recorded.

They did not look at existing guidelines for digitizing particular document and object types at the initiation stage, but did so later. For example, they looked at TASI, the Joint Information Systems Committee's (JISC) Technical Advisory Service for Images, and at the Visual Arts Data Service (VADS), which is part of the Arts & Humanities Data Service and is funded by JISC and the Arts and Humanities Research Board in the UK. They also talked to several people with imaging expertise. They did not identify a particular standard or a specific way of organizing things that would apply to the whole project. There were no guidelines that they rejected.

15.1.8 Target Audience and Evaluation

The primary intended audience for the digital deliverables was the general public. The secondary audience was the whole spectrum of teaching and learning, and third the subject specialists, such as archive users and conservation professionals. At the moment the website is used more by the general public.

Extensive evaluation of the targeted audience was carried out between the pilot and the main application. This was both qualitative and quantitative and included the educational users (for example, sessions with teachers about how the material from the project could be used in conjunction with National Curriculum areas) and the general public. Market research was carried out by a professional company.

They are currently looking into ways to cater for those with disabilities.

Where there are limitations on the use of the digital deliverables, these are stated clearly by disclaimers on the

website (e.g. use of text from the Stationary Office).

15.2 Project Management and Planning

They did not use external advice or consultancy on managing the digitization project, as there is sufficient expertise in-house. The National Monuments Record and English Heritage have considerable project management expertise and a long track record of managing digital records.

Eighteen new members of staff were hired for the project, and one project manager was re-deployed. They have adopted a very flexible approach based on team work and have cut a lot of red tape. The project has had an influence on the wider organisation in various ways. For example, the project has put more material on the web than any other part of the organization. Considerable knowledge has been gained on digital imaging and on web-enabling datasets, which has proved useful in other areas of the organization. Volunteer management was also observed with great interest by the rest of the organization, since it involves a different way of working. On a very simple practical level this meant, for example, that they had to design a different, much simpler expense form than their usual one, for the volunteers to fill in.

Project management procedures include: a Project Board, which meets at least quarterly; a Project Assurance Team and IT Assurance Team; and a User Group. The latter includes carefully selected members of staff to represent different parts of their audience, such as experts from architectural history or from communications media. This has proved very useful, even though its management has not always been easy. They started public access trials in June of 2001. They also have regular meetings with two HLF monitors, a lead monitor and an associate IT monitor. They use PRINCE II as a project management tool – this is probably one of the most PRINCE-managed projects in the organization. Although PRINCE can be overly involved in documentation, it is generally excellent for planning. Both the project manager and the IT manager are PRINCE-trained.

Digitization is outsourced. They have contracted Ewan Associates Limited, which they selected after an 18-month long, European-wide procurement process. The company sub-contracts different aspects of IT tasks related to the project, such as data capture and web design. The reason for selecting the option to outsource was that there were no resources in-house and they did not want to invest in the type of equipment and specialized staff that the medium and volume of this project required.

Each volunteer photographer receives an addressed pre-paid envelope and a form to fill in for every film they are given from the project. After they finish a film, they post it directly to the company doing the data capture who print the photographs and send them to the volunteers. The contractor scans the negatives, puts them on CDs and sends the digitized images to the project offices. All digitized images are checked for both photographic quality, metadata, and scan quality (pass/fail). If they fail the scan quality, they have to be re-scanned. After digitization and quality assurance of the images is complete, the negatives are also returned to the photographers. Because of the type of contract that they have, there is a set deadline, set budget, and set quality requirements.

They carried out a pilot study to provide them with information to assist project management and planning. This provided information for scheduling (for example, giving them a better idea of the different progress and cost of photography of urban, suburban, and rural areas), training needs, technical feasibility, technology forecasting (e.g. the web was very new when they started) and to some extent also for workflow analysis and piloting. User needs' analysis was carried out after the pilot and prior to the application being made to the HLF. After the pilot study, the design of the project was scaled up. Time and motion studies were part of the initial pilot, which provided them with information about the rate at which the material could be produced and the number of volunteers needed. The project team itself took photographs of buildings as a small pilot, which informed the process of providing guidance to the volunteers and was very useful. They use Gantt charts, staff charts in Excel, flow diagrams, resource leveling. They produce job descriptions (where there was some flexibility in adjusting job descriptions from the original HLF application) and an annual list of targets to be achieved by each member of the team as part of performance-related pay.

The external contractor uses Sony film scanners for digitizing the negatives. The project established a set of specifications for the quality of images that is acceptable and guidelines for scanning. Although they use grayscales and color charts, these are more suitable for controlled indoor conditions. It is very difficult to apply them in practice, because of the huge volume and variability in lighting conditions and exposure when the photographs were taken. They try to come up with a compromise of color balance correction applied to the images for acceptable quality, as they cannot color-balance each image individually.

15.3 Human Resources and Training

Apart from the volunteers, external contractors, and collaborates, there are 19 members of staff who work full-time on the project. These are:

Type of Staff	Number	% of time on the project
Director	1	2.5
Project Manager	1	100
Technical Manager	1	100
Image Selectors	8	100
Finance Officer	1	100
Office Manager	1	100
Office Support Staff (necessary for all HLF and internal paperwork as well as supporting the survey and communications work)	1	100
IT contract administrator	1	100
Communications Officer	1	100
Survey Coordinators	3	100

Only the project manager was re-deployed from another area of the organization. The background and profile of most staff is professional (financial, press, architectural, project management, IT and GIS, architectural history and archaeology). They design a training plan for each member of staff each year. The NMR is part of the Investor in People scheme (the UK initiative which encourages use of standards in training and other areas of

personnel development and regularly monitors organizations to see whether they still qualify to be part of the scheme).

Areas where training needs were identified included: managing and appraising staff; presentation skills; interviewing; architectural interpretation; use of post digitization and other software (e.g. Photoshop, DTP packages, MS project); handling of the media and dealing with customer complaints. They have a system of shared learning regarding architectural history, whereby members of staff give presentations to colleagues about their specialized area. All members of the project team are eligible for training. This was organized in a variety of ways: in-house, using project staff, external consultants, by attending external courses, with independent study, and by learning on the job. The training met the needs of the organization.

15.4 Project Life-cycle Processes and Procedures

15.4.1 Reproduction and Copyright

All copyright of the photographs rests with the volunteer photographers. The project signs a licensing agreement with each photographer to use their images in digital form. They will also have to draw up licensing agreements for third parties who want to take a copy of the digital images and use them in various ways. Users can download thumbnails and medium quality images and associated documents and use them for personal and research purposes, as long as it is not for commercial purposes. No electronic management systems are used to control copying.

15.4.2 Preservation/Conservation

This does not apply, as no analog material is kept.

15.4.3 Preparation and Sources

They use the cataloging and reference systems used by NMR for listed buildings (a custom-built Oracle database). They used a subset of the original information stored in the database for the digitization process. The contractors also record some quality data associated with the negatives (e.g. whether they are scratched or otherwise physically damaged). Very little material was rejected before scanning, only when the film was physically damaged or not exposed.

15.4.4 Metadata

They use the English Heritage & Royal Commission for Historical Monuments England (RCHME) Thesaurus of Monument Types, controlled vocabulary, and classification system. The metadata records information about the original object, the digital object, and technical and staffing details (e.g. who carried out the QA). Volunteers supply descriptive information, then some information is added by the image selector about the quality of the scan and of the photograph that is classified, using four grades. This means that the people who record metadata include volunteer photographers, the images selectors who are specialists in architectural history, and for some of the metadata information, the contracted digitizers. The catalog for digital deliverables is in electronic form on an intranet server. A section of it is available on the web. The original digital deliverables and the associated records are related on a single database.

15.5 Format, Resolution and Compression of Digitized materials

15.5.1 Images

The file formats used for images were JPEG for delivering and TIFF for capturing and preserving. They initially used Photo-CD before they realized its limitations for preservation. The resolution used is 2200dpi for capturing and preserving and 120*80; 360*240; 3072*2048 for delivering images. The internal bit depth is 12 bits per band and the output color depth is 8 bits per band. LZW (Lempel-Zif-Welch) compression is used by the contractor when copying the images on the CD. The original scans are retained in uncompressed form for preservation. JPEG is used for delivering. The aim of the compression is to improve access and to enhance usability.

Some post-scanning processing is carried out on the images. They use Photoshop for color corrections, removal of visible defects (e.g. scratches and dust). The original scans are not cropped but this may be done in the future for web delivery (e.g. to remove irrelevant features) using bespoke software. The average file sizes created are 18 MB at capturing and preserving stage (12-12 MB with LZW), and 10K (thumbnails) and 6MB (medium quality)

at delivery. The dynamic range of the contractor's equipment is specified at a minimum of 3.0 : 1.0.

Their experience showed that the digitization contractors were not accustomed to dealing with large volumes of material. In their case it was difficult to find someone who could apply consistently, the level of correction and guaranteed quality of output with the kind of varied input that the project produces. A large number of initially supplied images were rejected because of the presence of defects. The initial contractor may have assumed that some corrections and editing would be carried out by the individuals within the project. The current contractors are contracted to do 100% correction of images.

15.5.2 Quality Control

In terms of Quality Assurance, the way they worked (according to PRINCE guidelines) involved deciding what they wanted to produce at each stage, such as a handbook for the volunteer photographers, using a three month planning horizon, before defining what the acceptable quality for that would be and how they would check it. As quality was very important for the project, they spend a lot of time with the volunteers explaining about the expected quality, also including in the manual information about security, health and safety, example lists of entries which the photographers would have to produce, what constitutes a defining image and how to choose their views of the buildings, lenses. (Quality control procedures are also described above in the Project Management and Planning section.)

15.6 Delivery, Access and Use of Digital Deliverables

Users have free open access to both the catalog and the digital deliverables for browsing. Some of the material is available on the web. In October 2000 they put onto the prototype website the first 15,000 images which can be searched by type of building, date, county, historical figure or individual associated with the building, and by photographer. There will be terminals at the Swindon and London offices for the general public to use, with access to all the material, once the project has been completed. The web interface is a front-end to the database. They have taken a complex database and simplified it. All the fields that are indexed are metadata. They have no special hardware or software needs at the moment. They will be asking users to register if they want to receive added functionality in the future, and will also be developing e-commerce software. They monitor use via automatic data capture and web logs. The level of usage of the web pages is high. There were 480,000 images viewed in January 2001, a sharp increase from previous months related to press work on the project that was carried out in the US and Canada (on TV and radio). Most usage is from the English-speaking countries, but there have also been some surprises, such as Lithuania. Generally they find much useful information in the analysis of the statistics and the feedback for the users.

15.6.1 Dissemination and Publicity

In addition to announcements on the English Heritage website, press releases, a national press campaign when the project started, staggered local publicity during the survey periods, print and broadcast media coverage (such as the January 2001 one from the US), the project produces a free newsletter which comes out three times a year. They also participate at conferences and events (e.g. a demonstration at the Historic Buildings, Parks & Gardens Event 2000).

15.7 Evaluation, Funding and Long-term Sustainability

The project is funded at £3.9 million from the Heritage Lottery Fund, through the Millennium Festival Fund between 1999 and 2002. This was the single largest grant awarded through the Millennium Festival Fund. Additional support has come from a development grant from English Heritage itself. The total cost of the project is £4.4 million. There was also some private sector sponsorship with Kodak's donation of the photographic film used for the 1999 survey season.

The HLF have asked the project to provide a conservation plan for digital archiving, a risk assessment plan, a front-end evaluation report, a project management plan, cost models, workflow reports, a marketing policy, an education strategy, a post-project management plan, as well as plans for every stage of the funding and monthly highlight reports.

Within the remit of the project, no elements will be changed (apart from the file formats that had to be changed from Photo-CD to TIFF and updating the user interface) unless there is a specific need for them to do so. The NMR has a preservation strategy for both digital and paper records. The strategy is based on migration of data.

Quality control procedures in the life-cycle management are based on the adoption of open or *de facto* standards and other procedures carried out by the NMR and not by members of the project's team. The HLF contract specifies that they would need to keep the digital deliverables available for 25 years (for monitoring purposes). They plan to have them available in perpetuity.

15.8 General Comments, Lessons Learned, and the Future

As general comments, they emphasized the need for flexibility in digitization projects because the problems often do not emerge until the project is underway, but that planning sensibly can accommodate many of the problems. Projects must also allow sufficient time to plan effectively.

Particularly interesting aspects of this project include: the management of volunteers which seems to work very effectively and has the secondary output of raising awareness of the project and involving the wider community from the outset; the use of the PRINCE model for project management which demands considerable time from the project staff, but which has helped them in planning, coordinating, and carrying out effectively all the steps of the digitization process; the use of quality assurance procedures to monitor the quality of photography, scanning, linking with and entry of metadata and the use of the resource.

valid xhtml 1.1
abp~04/02

II. Project Planning

Introduction

Planning is the first and arguably the most important step in any digitization project. Lest this sound like a platitude, it is worth noting that far too many projects are undertaken without adequate thought to the activities involved, the staff required, or the technical exigencies of the work. The need for good planning may be self-evident, but in practice it is often difficult to anticipate all the areas in which forethought is essential. Good planning for any project—even for managers who have successfully completed previous projects—requires a large number of decisions on questions such as the following:

- What work needs to be done;
- How it will be done (according to which standards, specifications, best practices);
- Who should do the work (and where);
- How long the work will take;
- How much it will cost, both to "resource" the infrastructure and to do the content conversion;
- Where, after having answered all of these questions, one might obtain funding.

This kind of planning is one of the most intellectually challenging of the project tasks, and may well be time-consuming. There may also be pressure to hurry this step, from a desire to show visible progress or in response to institutional pressure. But an investment in this kind of planning will be amply repaid over the life of the project: in the quality of the products, in smooth workflow, in staff morale, and not least in the total project cost. The goal of this section is to sketch out the parts of the planning process and indicate the important decisions—assessing the resources needed to complete the project, the staffing and equipment required, the choice and role of metadata, and the overall project management—and how to go about making them effectively. The checklist below gives a brief inventory of the resources required to undertake a digitization project. Not all projects will require all the resources listed, but this list will show the range of needs you should anticipate.

Technology develops and changes so quickly that decisions like those listed above may seem almost impossible to make with any confidence. Information on the array of standards, specifications, equipment, skills, and techniques not only presents a daunting learning curve, but also a welter of detail that can be very difficult to track. For the project planner, however, it is not these details that really inform good decision-making. It is much less important to know what sampling rate a particular piece of equipment offers than to understand how sampling works and how it can affect the quality of digital conversion. These underlying principles apply more broadly and change more slowly. Most importantly, though, they represent the level at which good planning takes place; with this knowledge, the planner has the tools to bring together an expert group of staff and consultants and create an effective framework within which they can work. This *Guide* contains detailed, up-to-date information on best practices in a number of technical areas, but the *Guide's* greatest and most enduring value for the project planner is its presentation of the more fundamental issues and how they interrelate.

The *Guide's* introductory section has already addressed the first question on the list above: What work needs to be done? By emphasizing the identification of audience and of your own institutional location and goals, the introduction contextualizes this decision and reminds us to ask "Who needs this work? Who will benefit?" The further ramifications of this question are explored in Section III on selecting materials, which discusses how to assess your collections and set priorities for digitization, and in Section XII on user evaluation, which provides guidance on how to assess the needs of your audience and how this information can shape your digitization strategy. This is also the stage at which you should get the facts and make your decisions concerning rights management, without which you cannot proceed with digitization: you need to establish the intellectual property status of the materials you wish to digitize, and you also need to decide on your own strategy for managing the intellectual property you are about to create. Both of these issues are explored in depth in Section IV. And although the project's final product may seem impossibly remote at this stage, you need to consider how the results will be distributed: not only what technologies you will use, but also how you will control access and ensure that you reach your intended audience. Section X covers these issues in detail.

The question of how the work will be done—the specifications, standards, and procedures you need to establish—has many facets which are addressed at various points in the *Guide*. Foremost among these is the question of standards: by using standards-based approaches wherever possible, you increase the longevity, portability, and interoperability of your data. You need to be aware of the standards that apply to the kinds of digitization you are undertaking, and these are described in detail in the sections on digitizing text, images, and audio-visual materials. Given the complexity and breadth of most standards, though, you also need to be aware of the best practices that apply to your community. For instance, both documentary historians and linguistic researchers use the XML-based Text Encoding Initiative Guidelines to encode textual data, but each group uses the standard in different ways that serve their particular needs. While you are considering the specifications for your data, you should also think carefully about how to capture and represent the metadata you will need to administer your digital materials and enable them to be used effectively. The *Guide* includes an appendix on metadata which describes the various types and their uses. The relevant sections of the *Guide* also provide pointers to specific information on best practices for particular digitization communities.

The question of "how" also involves decisions about equipment. For the project planner, these questions are most usefully addressed not at the level of specific brands and models, but by thinking about the functionality you require and the tradeoffs you are willing to make (for instance, whether keeping costs low is more important to the project's overall success than achieving the highest possible capture standard). The sections on images and audio-visual materials discuss how to approach these decisions; more specific information on particular kinds of equipment can be found in the appendix on equipment. Finally, you need to establish an effective workflow for your project. At the highest level, this includes project management strategies, which are discussed later in this section, and quality assurance methods (discussed in Section VIII). But in addition you need to consider how you will store, manage, and track your digital objects, which is addressed in detail in Section XIII on digital asset management.

Staffing issues—who should do the work—are closely related to the points just mentioned, since your decisions about methods and procedures may be difficult to separate from the staff resources you actually have available. Few projects have the luxury of hiring all new staff to an arbitrary standard of skill and experience. Further on in this section we discuss human resources: how to construct job descriptions and identify skilled staff, and how to set up a management and advisory framework that allows your staff the autonomy to do their jobs effectively. In Section IX, Working With Others, we consider a range of collaborative and cooperative relationships that may expand your staffing options, including project consultants, vendor outsourcing, collaboration with other institutions, and internal cooperation.

Once you have worked through the issues sketched above, you will be in a position to assess the practical scope of the project: how long the work will take, and how much it will cost. Of all the questions addressed here, these may be the most vulnerable to change over time, as techniques and equipment improve and grow cheaper, and as quality expectations rise. Some guidance on cost estimation is offered later in this section, and also in the sections on specific digitization areas (Sections V, VI, and VII). You should make sure in researching costs to take into account all of the startup and infrastructural costs the project will incur-costs for initial planning, choosing data specifications, building or choosing tracking and documentation systems, training staff, and so forth-as well as the incremental cost of digitizing the materials themselves. This is also an opportunity to consider the scope of your investment and whether this infrastructure can be reused or adapted for further digitization projects once this project is completed.

Finding the funds to undertake the project is the final step, at least logically; a successful funding request will almost always require a thorough consideration of the issues just described. Even if you are fortunate enough to have funding already committed, going through this process will ensure that you spend your resources prudently and receive value for your investment. Funding sources and strategies are discussed later in this section, and also in Section XI on sustainability.

The checklist box below gives a condensed list of the resources you may need to undertake a digitization project. Although not all projects will need all of the resources listed, it gives a sense of the range and options.

Checklist Box:

Resources that you will need for a digitization project:

Personnel:	advisors project management staff rights specialists researchers editors authors digitizers catalogers technical support/development legal advisors	☐
Software:	operating systems applications: --> image manipulation --> metadata authoring --> database --> indexing and search engine --> web server utilities server systems network clients specialist applications/developments	☐
Storage devices:	local hard drives network storage servers optical devices (e.g. CD writers) magnetic devices (e.g. tape drives) controlled storage environment	☐
Network infrastructure:	cables routers switches network cards ports	☐
Consumables:	stationery utilities printer cartridges lamps (for capture devices/special lighting) storage and backup media	☐
Project management:	preparing bids recruitment publicity and dissemination creation of deliverable product specifications design of workflow supervision of staff quality assurance	☐

Resources within your institution

If you are working within an institution that has other digitization projects under way, an examination of the resources already available within your institution is a good starting point. Staff will know if their department or unit has capture devices available or workers with experience of digitization or cataloging. This is an easy first step towards building a resource inventory, although knowing that you have one flatbed scanner, a digital camera and suitable equipment for digitizing audio, as well as people who know how to use that equipment, is

not on its own sufficient. A thorough identification of internal resources involves checking that:

- equipment and software are of a sufficient specification to meet your requirements;
- workers who can operate the equipment are available and appropriately trained;
- technical support and maintenance are in place;
- capture devices are (or can be) directly connected to your storage area; and,
- access to equipment and staff suits your workflow requirements.

Clearly assessing the adequacy of these resources is predicated on other decisions, such as your workflow requirements; indeed, many of the planning areas discussed in this section are closely interdependent. It should also be apparent why the *Guide's* introductory section stressed early on that you need to define what you want to do and the audience or audiences you intend to reach (see Section I). A clear statement of objectives (preferably in a formal document that can be shared with staff), combined with the resource inventory, will enable you to assess the suitability of your local resources.

You will make this document an even more effective planning tool by adding information about equipment specification (e.g. computer processor speed, RAM, hard disk capacity) and the results of tests for suitability. Before you can conclude that you have suitable resources you must test them to make certain that they will meet the requirements of the project. The Example Box below, "Resource Inventory and Test", shows what a resource inventory and test for scanners might look like.

Example Box:

Resource Inventory and Test:

PCs and Scanners	Functional Requirements	Suitability Test Result
1 Pentium 3, 600 Mhz, 128 MB Ram	Must handle processing and manipulation of image files up to 50 MB	Needs more RAM
1 Pentium 4, 1 Ghz, 384 MB Ram		***Okay***
1 Agfa Arcus		***Okay***
1 Agfa DuoScan 1200		Transparency tray inadequate

Overall Conclusion: Upgrade one PC and replace one scanner

Most large institutions in the cultural heritage sector will have resources that may be useful to the project but would not necessarily need to be borrowed for the entire life of the project. There may be physical equipment, such as devices for digital capture, analog capture equipment (e.g. record, tape, CD and video players that can be used when converting from analog to digital), network storage devices, or handling equipment and controlled storage for analog material.

Human resources may be even more useful—expertise in digitization, text encoding, networks or web delivery can often be found in-house. Even those institutions yet to carry out any significant digitization will have cognate areas of expertise. These skilled individuals can be difficult to find, so tell your colleagues that you are planning a digitization project and have them consider which skills might be of value to you. For example, the skills, techniques and processes required by digital photography are identical in many areas to analog photography, and the same applies to image processing. Similarly, the standards and methods for creating metadata have their roots in the creation of bibliographic records, library catalogs or finding aids and museum collection

management systems. In addition to this, it is important to consider the project team and project management process here. Projects should establish a set of procedures for project management from the very start of any project, identifying goals and time scales as well as tasks and outcomes tied to the availability of specific staff and equipment.

It is much easier to identify potential facilities and expertise within the framework of an institutional digitization policy or corporate technology plan—follow the more detailed questions for your own resources as described above. If such a policy has not already been adopted, it will probably be beyond the scope of an individual project to initiate one. Nevertheless, informal inquiries can still be made relatively easily. Remember that apparently unrelated departments or projects may be useful. For example, a great deal of high-end digital imaging takes place in dental, medical, biological and life science departments. The Internal Resource Identification Question Box illustrates some of the common areas of expertise to be found within an institution.

- Is there an institutional digitization policy to adhere to?
- Who else in the institution has digitization projects underway?
- What experience can you use (e.g., photographic, equipment analysis, etc.)?

Question Box:

Internal Resource Identification:

	Institution Type		
Resource	Academic	Library	Museum/Gallery
Imaging	Medical Imaging / Media Services / Photographic Services / Library	Special Collections / Photographic Dept	Imaging / Publications Dept
Metadata	Library	Cataloging Finding Aids	Collection Management Finding Aids
Text Encoding	Literature / Language / Computing Science Depts. / Information Management / Library	Cataloging / Information Management Finding Aids Electronic Texts	Finding Aids / Information Management

External Resources

Identifying resources outside your immediate department, unit or institution can be a more difficult process. Success depends upon what type of institution you are, your strengths and limitations, the accessibility of the resources you are seeking, and whether there is scope for collaboration. Guidance from and access to the experience of others are likely to be readily available. The Link Box points you to national organizations that provide information to support digitization projects. Outsourcing can be another way to fill gaps in the resources available locally, by contracting with a vendor, hiring a consultant, or establishing a cooperative relationship with another institution. These options are discussed in greater detail in Section IX, Working with Others.

Link Box:

Links to National Organizations Offering Guidance

CLIR: Council on Library and Information Resources: "The projects and activities of CLIR are aimed at ensuring that information resources needed by scholars, students, and the general public are available for future generations." http://www.clir.org/

DLIB Forum: "The D-Lib Forum supports the community of researchers and developers working to create and apply the technologies leading to the global digital library." http://www.dlib.org/

LOC: Library of Congress: "The Library's mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations." http://www.loc.gov/

NINCH: National Initiative for a Network Cultural Heritage: "A coalition of arts, humanities and social science organizations created to assure leadership from the cultural community in the evolution of the digital environment." http://www.ninch.org/

RLG: Research Libraries Group: "The Research Libraries Group, Inc., is a not-for-profit membership corporation of universities, archives, historical societies, museums, and other institutions devoted to improving access to information that supports research and learning." http://www.rlg.org/rlg.html

PADI: "The National Library of Australia's Preserving Access to Digital Information initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access." http://www.nla.gov.au/padi/

AHDS: Arts and Humanities Data Service: "Create and preserve digital collections in all areas of the arts and humanities." http://ahds.ac.uk/

HEDS: Higher Education Digitization Service: "The Service provides advice, consultancy and a complete production service for digitization and digital library development." http://heds.herts.ac.uk/

TASI: Technical Advisory Service for Images: "Advise and support the academic community on the digital creation, storage and delivery of image-related information." http://www.tasi.ac.uk/

Resource challenges

There are a number of challenges both in assessing and securing the resources required for the project. Projects that take place in large institutions frequently benefit from a significant amount of non-project-related investment. Such hidden benefits include local area networks, high bandwidth Internet connections, large capacity network-based storage devices, web servers, and technical expertise associated with maintaining and developing these facilities. This infrastructure provides the framework for the specific resources and skills a project needs, and without it many projects simply would never get off the ground. Although institutions are now trying to quantify this input, its actual value is difficult to establish, with the result that projects in well-resourced institutions are able to scale up more quickly but often under-represent the real costs that lie behind the their activities.

Equally, less well-resourced institutions and initiatives face an increasing challenge in matching the developments in presentation and delivery of digital resources that larger projects can provide. Frequently, the solution is for small and medium size institutions to develop collaborative projects. The Colorado Digitization Project (http://coloradodigital.coalliance.org/) provides a flagship example of how equipment, staff and expertise can be shared between large and small projects alike, enabling the digitization and delivery of resources that would not otherwise be possible.

Another challenge for digitization projects, large and small, lies in the area of human resources. Content creation is a burgeoning field and although many Internet businesses have failed, those companies such as Getty Images, Corbis, The Wall Street Journal and Reed Elsevier, which have adopted prudent content creation and marketing strategies, are showing steady growth. The finance, commerce, media and entertainment industries all recognize the value and benefits of digital assets, and this places a premium on skilled personnel. Furthermore, the development of staff with digitization skills related specifically to the humanities and cultural field has not kept pace with the growth in the number of digitization projects. Many projects report difficulties in recruiting and retaining staff. Few public sector projects can match the remuneration levels offered by the private sector, but there are strategies you can adopt that enhance your chances of meeting the human resources

challenge. These are outlined in the Human Resources Question Box.

Question Box:

Human Resources:

- Are there non-monetary factors that can be emphasized or enhanced? For example, will the project offer advantageous working conditions, training opportunities, or the possibility of gaining qualifications or accreditations?
- Are there aspects of the job that are more attractive than private sector equivalents (e.g. greater creativity, responsibility, freedom)?
- Can posts be combined or split to make most effective use of existing skills?
- Can you consider applicants from a non-humanities/cultural background, particularly for technical posts?
- Can any staff be re-deployed, temporarily transferred or re-trained from elsewhere in your institution?
- Can posts be shared or joint funded with other projects?
- Are you able to outsource any jobs?

Funding

Some project staff will be preoccupied with securing adequate financial resources to start, develop and sustain a project throughout its lifecycle. An accurate picture of the financial costs will help you to identify the financial pressure points and to estimate more accurately the overall costs of running the project. The sections below on skills, equipment, and project management will provide points to help you develop accurate project budgets. An accurate profile of project costs helps to minimize the financial unpredictability of the project and improves the probability that it will attract funding. Funding agencies remain attracted by the opportunities for funding initiatives in the heritage sector. The Link Box provides pointers to some major US funders.

Link Box:

Potential Funders of Digitization Projects:

- Andrew Mellon Foundation: The purpose of the Foundation is to "aid and promote such religious, charitable, scientific, literary, and educational purposes as may be in the furtherance of the public welfare or tend to promote the well-doing or well-being of mankind." http://www.mellon.org/awmf.html
- NEH: National Endowment for the Humanities, "an independent grant-making agency of the United States government dedicated to supporting research, education, and public programs in the humanities." http://www.neh.gov/
- The Getty: "The Getty Grant Program provides support to institutions and individuals throughout the world for projects that promote the understanding of art and its history and the conservation of cultural heritage." http://www.getty.edu/grants/
- IMLS: Institute of Museum and Library Services, "an independent federal agency that fosters leadership, innovation, and a lifetime of learning." http://www.imls.gov/grants/index.htm
- NHPRC: National Historical Publications and Records Commission, "supports a wide range of activities to preserve, publish, and encourage the use of documentary sources relating to the history of the United States." http://www.nara.gov/nhprc/

From the projects surveyed it is evident that most potential funders, particularly in the public sector, require applicants to provide a robust and auditable cost model. How this should be presented may vary from one funder to another, but it can be extremely useful to break down equipment and salary costs on a per unit or work package basis. Not only does it help the potential funders to make comparisons of unit costs between

projects within and across heritage sectors, but it also forces you to look at the process and scheduling of work in detail. The accuracy of these figures will be greatly improved by conducting a pilot study or by adopting a cost model from a previous project, even if it needs to be revised in light of the experience of the earlier project.

All the projects surveyed obtained their financial backing from a combination of institutional budgets, public grants, private donation or corporate sponsorship. None of the projects reported serious under-funding, although some found that the distribution of funds created an uneven cash flow, resulting in medium term planning problems. Similarly, none of the projects reported serious concerns about sustainability, even where the source of future funds was unclear. The general absence of plans for self-generating funds or of exit strategies supports this confident view that income would continue to materialize in the future. A number of projects have recognized that failing to adopt long-term financial planning is less than prudent. We recommend that time and support for securing further external funds are crucial as well as exploring the potential for self-generating income. Projects should develop an exit strategy that will secure the maintenance and accessibility of the digital material. These issues are discussed in more detail in Section XI on Sustainability.

Cost models

Determining the cost of digital content creation on a per unit basis is extremely problematic. Not only are there no comprehensive cost models available that cover all resource types but trying to apply such a model to the variety of institution types, financial arrangements, prevailing market conditions, nature and volume of material and the resolutions required would be problematic. Furthermore, the cost basis for creating, storing and delivering digital resources can be quite different and trying to establish a single cost per unit can disguise these differences or ignore them altogether. In spite of these problems it is possible to establish some bases for per unit cost.

At the simplest level a project can take the total funding required and divide it by the total number of units that they intend to digitize. For example total project funding of $300,000 divided by 40,000 units equals $7.5 per unit. However, such a figure can be extremely misleading. Firstly, there will be variation in per unit cost according to the type of material digitized. The creation of OCR text pages will differ from reflective color still images, which will be different again from 16mm moving images or 78 rpm records. Even within material of the same broad type there will be variation. Black-and-white negatives are likely to be more expensive to scan than black-and-white prints, since tone reproduction needs to be set image-by-image in the former case, while the same settings can be applied to a group of photographic prints. Even if a project is dealing with material of a uniform medium and size, variations can occur that impact on unit costs. A collection of bound, legal-size books may have volumes that cannot be opened beyond a certain degree for conservation reasons. This may require a different capture technique, for example capturing pages from above rather than inverted. Some volumes may have details that demand a higher capture resolution than the rest of the collection, while others may require curatorial intervention to prepare them for digitization. The extent to which projects need to take account of such details will vary but at the very least different material types should be distinguished as well as same-type materials that require different capture techniques.

The cost items that go to make up a per unit calculation also require consideration. Should pre-digitization conservation work, handling time, programmers and management staff be included in addition to capture equipment and staff? In practice, projects need to do both. This is best achieved by calculating the costs directly related to capture on a per unit basis, which facilitates comparison and cost effectiveness for different techniques. Non-capture-related items could then be added to provide a total project cost and a second per unit calculation could be carried out if required. The list box below provides an indication of how these different factors can be differentiated. It is common practice to calculate costs for audio-visual material on a per minute basis.

List Box:

Capture Cost Factors:

(per unit for a single media type with uniform capture techniques and settings). It is important to note that the digitization capture costs are actually the least costly of the whole process.

- Handling time (from the shelf to point of capture and return) as a percentage of total salary costs on a

daily basis
- Pre-digitization conservation work (this should only be applied for those items that have required it)
- Capture time (from set-up to naming and saving) provided as a percentage of the capture operators total salary costs on a daily basis
- Cataloging/Metadata (required for digitization and/or created at capture stage) as a percentage of total salary costs
- Hardware cost per item
- Quality Assurance time as a percentage of salary cost
- Software cost per item (both hardware and software costs should be on the basis of the depreciation of equipment or projected replacement cost, rather than the total cost of hardware and software)
- Hardware maintenance
- Technical support time (proportion of total salary or contract cost related to capture)
- Project Management time (proportion of total salary related to capture)
- Training (directly related to capture)

Non-Capture Cost Factors:

- Additional Project Management salary
- Web Programmer's salary
- Educational Officer's salary (or other additional project staff)
- Cataloging/Metadata (post capture creation) % of total salary costs
- Additional technical support salary
- Additional hardware and software costs
- Consumables (including any storage media such as CDs or DATs)
- Travel and per diem
- Training (non-capture related)
- Storage costs (based on total maintained cost for the gigabytes required)

Some sites with detailed information on costing are listed below.

Key Sites with resources on costings:

- Research Libraries Group: http://www.rlg.org
- Library of Congress: http://www.loc.gov
- Online Computer Library Center: http://www.oclc.org/home/

Human Resources

A project's long-term success depends on the accurate assessment of the required human resources, and producing a map of available and unavailable skills is a valuable starting point. Institutions vary in their areas of expertise and different types of project require different skills. Nevertheless, from the projects that we surveyed it has proved possible to develop a basic template of the people and skills required in realizing a digitization project. The requirements can be scaled according to the size of the project envisaged.

Job descriptions, performance indicators, training

Comprehensive job descriptions are indispensable, regardless of the project or institution. While job descriptions are not always required by the host institution, employment law often demands them. Funders are increasingly expressing an interest in viewing job descriptions as part of the application process as this provides them with a richer overview of the project. It is worthwhile developing an outline of job descriptions before the project reaches the recruitment stage. This is useful to determine the delegation of work, how jobs interrelate, which posts can be tailored to existing skills and which can be identified for external recruitment or outsourcing. A

useful process for developing accurate job descriptions is to set out a list of all the tasks required for a post and then rank them from highest to lowest priority or into essential, desirable and non-essential categories. Next, compile a corresponding list linking these tasks to the skills required, including any particular knowledge or qualification. Alongside this, compose a description of the experience or background required for these skills. Finally, review the original tasks and their priority to ensure that a realistic and coherent job description is produced. A resource which has been developed by the Association for Computers and the Humanities is a database of jobs in this field—it may be consulted by projects for guidance in drafting job descriptions, and can also be used to publicize new jobs to a focused audience of candidates. See http://www.ach.org/jobs/ for more information.

Example Box:

Sample Job Description

Job title: Digital Library Research Assistant

The Digital Library Research Assistant will play an integral role in the university's digital library projects, the goal of which is to bring a wide range of source materials to as large an audience as possible. The DLRA has responsibility for overseeing initial scanning and data capture, creating and reviewing metadata, and performing quality assurance checks. With other project members, collaborates on project publications and research.

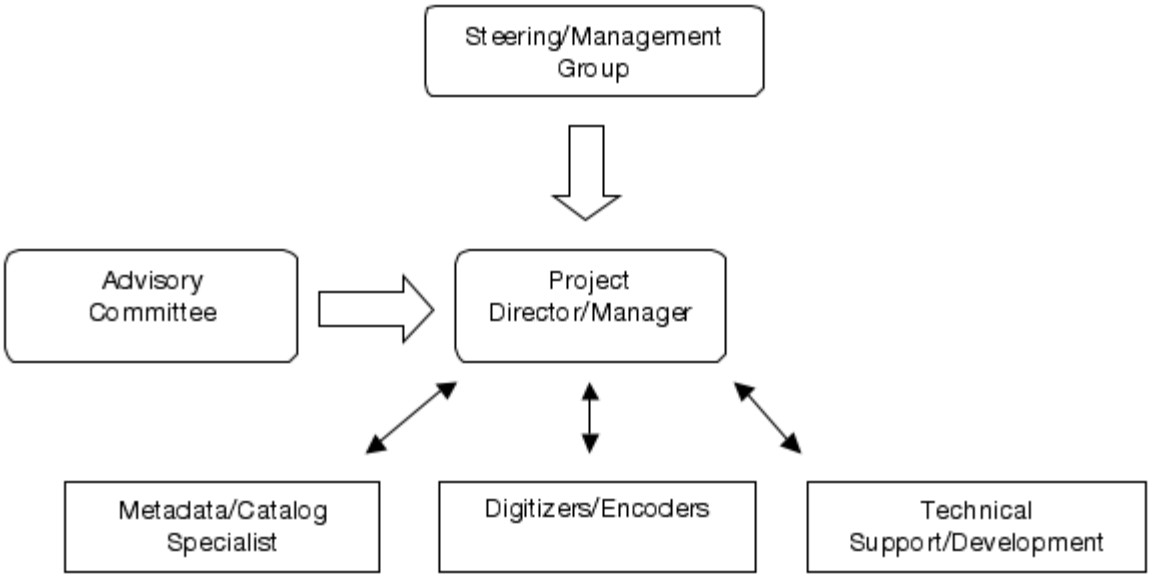
Job requirements: Bachelor's degree and one to three years' experience; basic computational skills, and expertise in at least one area of the humanities. Advanced degree and three to five years experience preferred. Familiarity with relevant encoding and metadata standards, including SGML/XML, METS and Dublin Core, is highly desirable. Must be a self-directed team worker with strong motivation and the ability to take initiative. Needs good communications skills (oral and written) and willingness to work collaboratively.

The use of performance indicators appears to be on the increase. They can have a positive impact, not least by providing a way of formally identifying training requirements. While most projects assess training needs on the job as an informal exercise, formal methods encourage appropriate training solutions to be planned and resourced in advance.

There is a close interplay between performance indicators, job descriptions and training assessments. The job description is very useful in developing meaningful performance indicators. Indeed, a useful starting point for performance review is to evaluate current tasks against those set out in the job description, highlighting whether the original job description was unrealistic, whether workloads need to be re-evaluated in the light of practical experience, or whether a skills shortfall needs to be addressed. The aim of addressing training requirements is to ensure that future tasks can be achieved and that the project will not encounter a skill shortage.

Managing the skills base

It is vital to ensure that a project be able draw on the right balance of skills. The challenge is to determine the skills of individuals and how they can most effectively contribute to the project. The key to successful delivery of projects is management. The diagram below incorporates elements from all of the projects surveyed, from the smallest to the largest, and illustrates the general structure that may be used to manage the project's skills base.



The steering group functions as an executive board and includes all constituents who are directly involved in the project, even if not employed by it, such as curators, archivists, subject specialists and education officers. In practice it is common for the steering group to be an existing committee within an institution.

The advisory committee is a broader-based group, providing general advice on the project's focus and direction. Members usually include the steering group with additional appointments from external organizations bringing particular areas of expertise, such as evaluation, to the initiative. There may be more than one advisory committee, or the advisory committee may be broken down into sub-committees each of which supplies more focused technical, academic or editorial decision-making support. This is the case with the Perseus Project at Tufts University, which has separate Technical and Academic Advisory Boards as well as a Steering Group to provide general project management. (Read Interview 28.2 for details on this arrangement)

It is essential to have a single project manager who is employed by the project, with responsibility for its daily management. In most cases the project manager provides the necessary project management experience, supplemented by internal or external advice. An institution needs to assign both accountability and authority to the project manager position, so that the process is not bogged down by myriad interactions with the advisory group or groups to deal with daily operations. In content creation projects it is unusual to employ external consultants to handle project management.

What skills are required?

There are four main areas, which will require staff with identifiable skills. These skill areas may be provided within a single project, dispersed across a collaborative project, or outsourced.

- Conservation: A crucial aspect of any digitization initiative will be a conservation assessment of the analog materials. Under some conditions this may show that before some material can be digitized it will require conservation intervention.
- Digitization/Encoding: This can involve digital imaging, keyboarding, OCR, character or full-text encoding, or a combination of these. In some projects it may also include conservation intervention in the analog material.
- Metadata/Cataloging: The creation of metadata records for the digital material. This work may also involve cataloging the analog material or searching for information to enhance the metadata record where it is absent from the analog version.
- Technical Development/Support: This falls into two distinct areas: the creation or implementation of specific IT solutions for creating, managing or delivering the digital material, and the provision of IT support for project hardware and software. This latter area includes desktop applications, network services, and capture devices.

In smaller projects staff may carry out tasks in more than one area: for example, the digitizer may also undertake technical development, or the project manager may take on metadata creation. In larger projects, such as SHOAH or the Genealogical Society of Utah, the duties of staff are so extensive that this is not feasible.

Project managers will have to decide whether to hire new staff with the required skills or to re-deploy existing staff from other areas of the institution. We found that many projects prefer the former, with two notable exceptions. First, there is a discernable trend for photographers to be employed for high-end digitization work. Projects have found that better-quality images are produced through training a photographer in digitization rather than trying to equip a digitizer with photographic skills. The second exception is the tendency to re-deploy or train existing cataloging staff in metadata creation. This is a logical progression for staff who will already have considerable experience in creating bibliographic records, collection management records, finding aids or catalogs, frequently in an electronic form such as MARC.

Another decision concerns background skills. With the exception of some technical posts, we noted a clear preference for staff with arts, humanities, library, museum or gallery backgrounds, or at least some experience or interest in the subject area of the collection. There may sometimes be advantages in not having such a specialization. For keyed-in text transcription, staff without subject knowledge are more likely to enter exactly what is on the page rather than interpret the contents and enter what they think is in text. On the other hand, subject knowledge can be exceptionally useful in gauging what areas of the content should be focused upon, deciphering difficult materials, or recognizing how areas of the content should be marked up.

When you are trying to find staff with appropriate skills, remember that some projects have benefited from using student labor and volunteers. The ability to draw on student labor represents a significant benefit for university-based projects. Projects such as those based at the University of Virginia Library have been able to build large and diverse digital collections because they are able to draw upon a pool of skilled, motivated and affordable labor. Projects that recruit student labor have invested considerably in training, adopted flexible working practices and tailored the work around the students' educational commitments. This approach has the added benefit of equipping students with the skill set required for future work, adding to the pool of available staff.

Volunteers often provide a similar pool of skills and projects such as the Genealogical Society of Utah have made effective use of this resource. They have found it both necessary and beneficial to invest in appropriate training for the volunteers. Such training should be factored into the project resource plans. In large-scale initiatives, volunteer management and training may become a significant part of the project itself.

The Link Box below provides links to sites that support skills development in digital representation.

Link Box:

An increasing number of organizations are offering training in digitization, which generally proves cheaper and far more useful than commercial training courses:

- Archive Builders: Document Imaging and Document Management. http://www.ArchiveBuilders.com
- Cornell University Library, Department of Preservation and Conservation: Moving Theory into Practice: Digital Imaging for Libraries and Archives. "This workshop is intended for librarians, archivists, curators, administrators, technologists, and others who are contemplating or implementing digital imaging programs." http://www.library.cornell.edu/preservation/workshop/
- HATII: Humanities Advanced Technology and Information Institute. Digitization Summer School: "The course will examine the advantages of developing digital collections of heritage materials, as well as investigate issues involved in creating, curating, and managing access to such collections." http://www.hatii.arts.gla.ac.uk/
- Humanities Computing Unit in Oxford: Summer Seminars covering a range of topics: http://www.hcu.ox.ac.uk/
- School for Scanning - North East Document Conservation Center: Creating, Managing and Preserving Digital Assets. http://www.nedcc.org
- Harvard Extension School, Museum Studies Certification Program. Muse E130: Information Technology for Museums. http://www.extension.harvard.edu/2001-02/courses/muse.shtml
- TASI: Technical Advisory Service on Images. Training courses "aimed at those involved in digitization projects, those who wish to capture images and those who wish to use digital images in teaching and research." http://www.tasi.ac.uk/training/training1.html

- UCLA/Getty course: Museum, Libraries and Archives: Summer Institute for Knowledge Sharing. http://skipper.gseis.ucla.edu/orgs/gettysi/html/summer.html
- University of Virginia Library: Digital Media Lab Services tutorials and short courses on digital image, video, and audio capture and editing. http://www.lib.virginia.edu/clemons/RMC/digilab-services.html

Equipment

Because our digitization capabilities are so strongly tied to—and limited by—the developing equipment technology, it is tempting to feel that the available technology should motivate our digitization strategies. However, on the contrary, it is vital to base equipment requirements on the characteristics of the collection/s to be digitized and on project needs, and not the other way around.

Although there are significant cost savings associated with outsourcing work to "offshore" production bureaus in Asia, the Far East, Mexico, etc, in cases where unique materials or special collections materials are to be digitized it is important that digitization should take place as close to the original as possible. Hence many projects will need to confront the complex questions of equipment specification and selection. A detailed discussion of matching material properties to hardware and capture settings can be found in Section VII on audio-visual materials. There is also further information on equipment choices in the appendix on equipment. At the moment we will focus on the basic differences in equipment and the technologies employed in order that the correct type of equipment resource can be procured for a project. Selecting the most appropriate equipment can be time consuming, but projects should not be deterred by the plethora of manufacturers and their competing claims. For example, the SCAN project (Scottish Archive Network) was initially unable to find a commercially available digital camera that exactly matched their requirements. Instead, they sourced a camera custom-made to their exact specification. This level of exactitude may be out of reach—and unnecessary—for most projects, but it is worth remembering that one need not be entirely constrained by what is commercially available.

Principles of digital data capture

Although there is a variety of capture devices for different applications, whether you are digitizing images, text, audio, video or 3D objects, the operating principles are the same. All digital capture devices take a sample of the analog source material to create a digital surrogate. This sample is made up of two elements: the sample rate and the sample depth. The sample rate describes how frequently readings are taken of the analog material. For example, in a digital image this would be the resolution, or the frequency per unit of area: the number of pixels per inch, expressed as pixels per inch (ppi) or dots per inch (dpi). An image captured at 600 ppi would have had 360,000 samples recorded per square inch. Similarly, for audio-visual materials the sample rate is the frequency per unit of time at which the source material is sampled. The sample depth is the amount of information recorded at each sampling point. For example, a sample depth of 24-bits would capture 8 bits for each of the three color channels (red, green and blue) at every sample point. For a more detailed explanation of sampling, see the appendix on digital data capture and Section VII on Audio-Visual Materials.

Selecting equipment

The medium, format, size, and fragility of the original material are among the primary factors affecting equipment choice. For text documents, flatbed scanners are suitable for single leaf, regular sized documents, provided the material does not go beyond the scanner's maximum imaging area (usually up to approximately US Letter size), or is put at risk by "sandwiching" it in the scanner. Large format flatbed scanners and sheet-feed scanners can handle single leaf, oversized documents. However, sheet-feed scanners put material at greater risk than flatbed scanners as the originals are pulled through a set of rollers. Drum scanners, whose imaging area is usually from 8" x 10" to 20" x 25", and digital cameras can also be used for oversize material, but they are an expensive option compared to flatbed scanners.

Bound pages that cannot be disbound, and pages in bindings that cannot open fully to 180 degrees require flatbed scanners with a right angle, prism, or overhead capture array. Digital cameras, with appropriate easels, book rests and weights are a versatile option for bound material. Camera beds or mounts, lighting, lenses, and filters all add to the cost and complication but make digital cameras more versatile tools for capturing manuscripts, bound volumes, original works of art, prints, out-size material and artifacts.

To achieve the highest quality scans of transparent media (e.g. 35mm slides and negatives, 6x4 and large format transparencies and microfilm) specialist equipment such as slide and film scanners, microfilm scanners or drum scanners should be used. Some flatbed scanners, with a dual light source, can handle transparent media though they often lack the dynamic range comparable to that supported by transparency scanners. However, you will not achieve as high a quality image as you would with a dedicated film or slide scanner. These have an inherently higher resolution, appropriate for the small size of the original, hold the transparencies more closely and securely, and frequently have negative color compensation to correct color casts for different types of film.

Audio and moving image materials present their own problems for digital capture. Not only is there a variety of source formats, including wax cylinders, 33, 45 and 78 rpm records, 8-track and cassette tapes, two-inch and VHS video in PAL and NTSC formats, but it is often very difficult to obtain access to analog devices for playback and linkage is difficult.

Definition Box:

Audio-Visual Facilities:

- Audio capture card required for sound material or video capture card required for moving images
- Source devices, such as 78 rpm record players and tape players.
- Mechanism for connecting these devices digitization equipment
- Intermediary device, such as a DAT (capable of handling ASEBU and SPDIF digital audio) machine

Link Box:

There are a number of audio and video digitization projects that are just getting started:

RAI: http://www.rai.it/portale

BRAVA: Broadcast Restoration of Archives through Video Analysis
http://www.ina.fr/recherche/projets/encours/brava/

COLLATE: Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material
http://www.collate.de/index.htm

PRESTO: Preservation Technology for European Broadcast Archives http://presto.joanneum.ac.at/index.asp

AMICITIA: Asset Management Integration of Cultural Heritage In The Interchange between Archives
http://www.amicitia-project.de/ami_home.html

The 3D representation of objects, from coins to buildings, is at the forefront of current digitization developments. At present the technology can be divided into two broad categories. The first, and simplest, is to create a moving image of an object. This is achieved by moving a digital camera around the object, or rotating the object in front of a fixed camera, while taking a series of still images. These images are then compiled to create a moving image of the object. The most common format for this is QuickTime VR. This is a reliable technology that requires a digital camera and mount or turntable. However, it does not provide a true 3D representation of the object because while only two planes are captured and displayed, it still represents 3D objects using two spatial planes. The viewer cannot manipulate the object, and the views provided are fixed and pre-determined.

Creating a true 3D representation of an object requires that the dimensions and features of the object be modeled. That is, the three dimensions of the object are represented in the computer as a set of coordinates. Attached to this "frame" are the textures of the object to provide the surface details. At present most 3D imaging technology remains in the sphere of industry. The technologies used to capture coordinates, render the model, and interact with the 3D representation (such as haptic feedback systems that allow one to "touch" the object, or 3D printing to create facsimiles) are often quite costly and require a relatively enormous amount of computing

processor power compared to the average desktop computer (in 2002). As such, 3D modeling devices remain application-specific, for example body imaging, prototyping or CAD/CAM applications. However, it was not long ago that digital imaging was the sole preserve of medical applications. During the next ten years we should see increasingly cost-effective and user-friendly devices that will bring 3D modeling into the mainstream.

Definition Box:

Virtual Reality:

Virtual reality can be described as an interactive, self directed, multi-sensory, computer generated experience which gives the user an illusion of participating in a three dimensional environment, even if a synthetic one. For cultural and heritage institutions, this may mean using virtual reality to create virtual representations of three dimensional objects in their collections or to create representations of environments, such as an Egyptian tomb, an ancient Persian palace, a historic Greek theatre or an ancient landscape. These three-dimensional objects could range from coins, vases, and sculptures to representations of whole rooms of collections.

Metadata

Metadata is an indispensable part of any responsible digitization program, and considerable attention has been paid to the definition of high-quality metadata standards for various purposes. (The appendix on metadata provides more detail on different types of metadata, and on specific metadata schemes and their uses.)The availability of accurate metadata is as important as the digital surrogates themselves for accessibility, usability and effective asset management. In many instances institutions will already have substantial metadata about the analog object (for instance, catalog records) much of which can be applied to the digital object. The project will be able to reduce its metadata creation costs by building on existing metadata. When selecting material for digitization you may wish to give priority to material for which partial metadata already exists.

It is crucial to remember to determine the status of the existing metadata, when you are assessing resource requirements. In an ideal world the existing catalog or finding aid would be complete and up to date. However, many libraries, archives and museums have a backlog of cataloging work, and part of a collection selected for digitization could fall into this category. Therefore, it may be necessary to devote time to locating missing information for your metadata records. You must then decide whether to seek information just for those fields required for the metadata, or to update the original catalog record in its entirety. Digitization provides an economical opportunity for institutions to expand their metadata, so consider the possibility of seeking extra funds or devoting more resources to this activity. Some of the new elements required for the metadata record of the digital object can be generated automatically: for instance, automatic metadata creation is a feature of much high-end digital camera software and of some OCR systems. Alternatively, a project may need to develop its own system, and can greatly improve the efficiency and accuracy of technical metadata. There is a general dearth of metadata tools, which poses a problem for the efficient creation and management of metadata for many projects. There is therefore likely to be a significant element of manual work, whether this lies in adding digital objects to existing electronic catalogs, creating records for web-based delivery such as Dublin Core, or implementing encoded metadata schemes such as EAD. Creating a metadata record will usually take as long as creating the digital surrogate and if detailed encoding schemes such as Encoded Archival Description or Text Encoding Initiative are used, this process can be considerably longer.

METADATA RESOURCES:

GENERAL METADATA RESOURCES

1. Canadian Heritage Information Network Standards Page: http://www.chin.gc.ca/English/Standards/metadata_intro.html
2. J. Paul Getty Trust, Introduction to Metadata: http://www.getty.edu/research/institute/standards/intrometadata/

3. Extensible Markup Language: http://www.w3.org/XML/
4. International Federation of Library Associations and institutions. Digital Libraries: Metadata Resources: http://www.ifla.org/II/metadata.htm
5. Text Encoding Initiative: http://www.tei-c.org
6. Metadata Encoding and Transmission Standard (METS): http://www.loc.gov/standards/mets/

METADATA MENTIONED ELSEWHERE IN THE GUIDE

1. Section III: Selecting Materials: Metadata & Interoperability. The Dublin Core metadata initiative http://dublincore.org/
2. Section IV: Rights Management: Technologies for Copyright Management and Protection.
    - The Open Digital Rights Language Initiative (ODRL): http://odrl.net/
    - Digital Object Identifier (DOI): http://www.doi.org
3. Section V: Digitization and Encoding of Text - Text markup schema. Text Encoding Initiative (TEI): http://www.tei-c.org
4. Section VI: Images
    - Descriptive:
        - Library of Congress Subject Headings (LCSH): http://lcweb.loc.gov/cds/lcsh.html#lcsh20
        - Categories for the Description of Works of Art (CDWA): http://www.getty.edu/research/institute/standards/cdwa/
        - Art and Architecture Thesaurus (AAT) http://www.getty.edu/research/tools/vocabulary/aat/about.html
        - VRA Core Categories: http://www.vraweb.org/vracore3.htm
        - Dublin Core Metadata Element Set: http://dublincore.org/documents/dces/
    - Structural:
        - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
        - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
    - Administrative:
        - A Web Hub for Developing Administrative Metadata for Electronic Resource Management http://www.library.cornell.edu/cts/elicensestudy/
        - Digital Library Federation, "Structural, technical, and administrative metadata standards. A discussion document:" http://www.diglib.org/standards/stamdframe.htm
5. Section VII: Audio and Video Capture and Management
    - Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
    - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
    - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
    - MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
    - Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
    - Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
    - Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
    - Library of Congress's Digital Audio-Visual Extensions to METS Standard http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
    - Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html
6. Section VIII: Quality Control and Assurance: Importance of Quality Control and Assurance of Metadata
7. Section X: Distribution:
   Metadata Harvesting
    - Clifford Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/mhp.html
    - Donald Waters, "The Metadata Harvesting Initiative of the Mellon Foundation," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/waters.html
    - The OAI MHP protocol: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
    - MHP tutorial: http://library.cern.ch/HEPLW/4/papers/3/
    - CIMI Working Group: http://www.cimi.org/wg/metadata/
    - CLIR Metadata harvesting project: http://www.clir.org/activities/details/metadata-docs.html
    - DLF and Metadata Harvesting: http://www.diglib.org/architectures/mdharvest.htm
    - University of Illinois at Urbana-Champaign Metadata Harvesting services: http://oai.grainger.uiuc.edu/

8. Section XIII: Digital Asset Management: "Metadata definition and management"
9. Section XIV: Preservation:
   Institutional Approaches
   - METS (Metadata Encoding and Transmission Standard) http://www.loc.gov/standards/mets/
   - UK Cedars project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
   - Australian Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metadata: http://www.nla.gov.au/padi/topics/32.html
   - NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
   - OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
   - Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf

Project Management

Many different approaches to managing projects are possible. While we found little evidence of the conscious adoption of a project management model, such as PRINCE 2 (http://www.kay-uk.com/prince/princepm.htm), most projects implemented many of the key features of successful project management. As understanding of digitization becomes more commonplace it may not be necessary to "hot house" prototype projects in the manner that many early projects experienced. However, it should also be recognized that integrating existing projects into host institutions often adds a layer of bureaucracy.

The Genealogical Society of Utah provides a good example of a comprehensive project management model. Each imaging project undertaken follows six stages:

1. Negotiation and project administration
2. Capture Convert Acquire
3. Image and metadata processing
4. Storage and preservation
5. Indexing and cataloging
6. Access and distribution

All projects will need to consider these six areas in setting up their own project management systems.

You do not necessarily need to adopt all the activities of a project management methodology; rather you need to scale the method to the needs of your project. The whole process should be determined by the project's objectives and rationale for creating the digital deliverable. Each process should be defined, together with the specific objectives to be achieved and activities to be carried out. The various roles and responsibilities should be detailed (defining job descriptions and breaking finances down aid in this — see above) and adapted to the size and complexity of the project. This should enable the efficient control of resources and facilitate regular progress monitoring. Regular reviews should be used to ensure that the project's objectives, which may change during the project lifecycle, are being met. Whatever project management method is adopted, it should provide a common framework and delineate milestones for all elements of the project.

In summary, your project management methodology should make possible:

- The use of pilot projects and feasibility studies to shape the overall scheme of activity
- Controlled and organized stages
- The establishment of a project plan with milestones
- Regular reviews of progress against plan and against objectives
- Control of any deviations from the plan
- The involvement of all constituents at the right time and place during the project
- Good communication channels between all constituents in the project and the sponsoring institution/s

Other key features are the need for one project manager to have ultimate responsibility and for the project advisory group to provide management quality control and assurance. In distributed projects, site managers are recommended in addition to an overall project manager. Most projects have relied on internal project management expertise, supplemented by external advice. Although many projects started as relatively autonomous there is a clear trend for project management structures and the project organization to be

integrated into the host institution's structure. This may be a natural progression for projects as they mature, but new projects may consider whether they should adopt it immediately.

Work flow and costings

While few of the projects interviewed carried out benchmarking tests most had conducted pilot studies. These were undertaken for a variety of reasons:

- Technical feasibility
- Technical forecasting
- Workflow analysis
- Training needs

When considering technical forecasting or prototyping, particularly in relation to costs, remember that there may be no corresponding benefit, and if there is a benefit it will vary for different types of content. Few projects in the humanities and cultural sector charge users for the digital deliverables. As such the cost/benefit may simply be realized by the ability of the project to amortize the depreciation on the equipment. A new high-resolution camera may pay dividends for fine textual or line art material, but not so for color images. Similarly, a device that enables the digitization of material that previously could not be captured, such as a 3D modeler, may not make financial sense if a project has to build in a profit or depreciation margin. However, if the device makes an important collection more widely available, the public access benefit may outweigh the financial costs.

Where any form of pilot study is undertaken it is important to build this into the project design and development cycle. For example, the University of Virginia Library's Special Collections department delineates its project work as intricately as possible before extrapolating its workflow and costings. This has given the project reliable data to forecast costs, but there are some areas where measurement has proved inaccurate, such as network transfer rates. The UVA Special Collections department also has a scheduling calendar tied to a tracking database to generate quality control and assurance checks and back-ups. In this respect it is typical of the projects surveyed which all use flowcharts, spreadsheets or Gantt charts to plan and monitor their workflow and costs.

If you are considering using a cost model (see above), it is important to include all the relevant costs, not just the obvious items such as equipment and staff time. You will also need to decide on what basis to evaluate — for example, costs per unit to be digitized or costs per hour. The table below provides a checklist of the factors that should be built into a cost model.

Finally, one further area to be aware of as you develop your cost estimates is digital asset management. In digitizing an image collection, for instance, you may well be generating a number of different kinds of digital objects-archival masters, delivery masters, thumbnails and other deliverables-which in turn will require storage, tracking, documentation, and upkeep. This process may require a significant commitment of resources and will need to be planned carefully. Section XIII covers digital asset management in detail.

Cost Model Factors	
Equipment	Purchase Maintenance Repair
Software	Purchase Upgrades
Staff	Salary (including benefits and insurance,) Training Recruitment Travel & subsistence
	Heat Light

Utilities	Water Phone Postage
Building	Rates Maintenance Upgrading/expansion

Analysis Box:

Costs of Digitization Programs

There is little information available about costs and this is an area where future work is necessary. In a rare exception, Steven Puglia analyzed the costs of digitization programs, in particular from the Library of Congress Ameritech competition and the National Archives and Records Administration's Electronic Access Report. The costs discussed are mostly *projected* and *estimated* costs, a problem discussed in the conclusion, which suggests that further studies are necessary. After an initial discussion on general costs of projects—it appears that on average, a third of the costs incurred by projects is the digital conversion, slightly less than a third is metadata creation, and slightly more than a third is made up of administrative and quality assurance tasks—the emphasis turns towards long term maintenance costs. The author suggests that these are not often taken account of with the project costs.

Three types of maintenance of digital objects are considered, each with mounting costs in relation to the initial costs per image:

- The first projection is an example taken from the maintenance of NARA objects. This involves only keeping one copy of master files and offering thumbnails and access online. The cost of this basic maintenance is in the region of 14-38% of the initial imaging cost.
- The second projection comes from Cornell report on computer output microfilm. Applying this to NARA, the cost would be 55-60% of the initial cost per image.
- The third projection looks at outsourcing to a commercial firm. This would cost 275-796% of initial cost. But it must be remembered that the other two costs do not include major IT infrastructure costs and thus are false figures, whereas the private firm costing will include this.

In conclusion, it is suggested that digital imaging may not be the best approach for long-term retention of information. Institutions can only justify retention if the images are used. Analog retention is the best way of holding materials in the long-term. In addition, it would be instructive to use figures from final project costs and also examine costs per person and production per person.

For the full report see: Steven Puglia, 'The Costs of Digital Imaging Projects', *RLG DigiNews*, October 15 1999, Vol. 3 No. 5. http://www.rlg.org/preserv/diginews/diginews3-5.html

Conclusion

At the start of any project, project planning feels like a way to exert control, eliminate risk, and guarantee a successful outcome. Certainly without good planning, the likelihood of failure and inefficiency is much greater. But you can be a better project planner by recognizing that the goal is not to eliminate risk but to prepare for it—not to control every variable but to create a project framework within which your team's response to the unforeseen will be resourceful and effective. In the technology domain, change and unpredictability are facts of life, and often represent opportunities rather than disasters for a well-planned project. Your planning goal should be to create a flexible, adaptable system whose staff and procedures can accommodate change. Your aim as a project leader should be to distinguish between what is essential—the central project objectives, the strategic components that will ensure long-term viability—and what is merely instrumental detail.

valid xhtml 1.1

XIV. Preservation

Introduction

The projects and programs interviewed for this *Guide* all acknowledge that digital materials are at risk of loss, both through physical damage to the storage medium and through technological changes that render the data unreadable. Although none of the projects we interviewed identified digitization as a chosen method for preserving analog material, digitization is starting to supplant microfilm as a preservation medium and hence the urgency of providing long-term preservation of digital materials becomes of paramount importance. For museums and institutions with two- and three-dimensional objects, moving images, and audio, the need is even greater, since microfilm has never been even a minimally adequate preservation medium for these. Overall, there is an urgent need for significant research and the development of affordable strategies for trustworthy digital preservation services. This *Guide* does not aim to cover the full range of technical and scientific developments in storage media and methodology, but explores some of the issues involved in preservation in order to explain the practical steps you can take to store and preserve your digital materials.

In this *Guide* we define 'preservation' as the actions taken to preserve digital objects, whether they are created from analog collections or are 'born digital' objects such as original digital art, and digital audio or video that are not re-recorded from analog originals. This domain includes two important and overlapping areas: first, the goal of presenting the content of analog items via digital reformatting; and second, preserving digital content regardless of its original form. 'Conservation' of the original analog objects—generally the domain of conservators, museum curators, librarians, and collection managers—is somewhat different, as it involves the care and repair required to keep the physical object in existence. However, there is some overlap between the two: objects are often digitized to help conserve the original object by reducing wear and tear on fragile and vulnerable materials, while still allowing access to them.

Preservation of digital objects can be thought of as long term responsibility to the digital file — responsibility to ensure that the file retains the information it had when it was created. Keeping the content 'alive' over time entails much more than simply keeping the files intact. Issues such as metadata must be understood as well. These issues are explored more fully in the rest of the section. There are two distinct areas of intention here, although in practice they often merge: preserving the digital surrogate, so that its content remains accessible and usable over the long term; and using digitization as a method of preserving objects which are valuable, fragile or transient. For example:

- Access to digital surrogates replaces direct use of rare and fragile originals, aiding their conservation.
- Some projects use digitization to 'capture' objects that are discarded after the capture process, such as newspapers.

Therefore the digital surrogate is a form of preservation (although not a substitute for any other form of preservation) and must itself be preserved to ensure future access and use (see Section XI on sustainability). The project should consider the two areas: digital reformatting and preserving content in digital form.

Why preserve?

Digital preservation is an essential aspect of all digital projects: it is imperative that the digital objects created remain accessible for as long as possible both to intended users and the wider community. Decisions made throughout the digitization process will affect the long-term preservation of the digital content. While the processes involved in digital preservation are the same whether for digital surrogates or for born-digital objects, remember that born-digital objects are the only version of an artifact or object so are, in effect, the preservation copy. All these factors must be incorporated in your digital preservation strategies.

CEDARS (http://www.leeds.ac.uk/cedars/) has described the digital preservation process as follows:

> Digital preservation is a process by which digital data is preserved in digital form in order to ensure the usability, durability and intellectual integrity of the information contained therein. A more precise definition is: the storage, maintenance, and accessibility of a digital object over the long term, usually as a consequence of applying one or more digital preservation strategies. These strategies may include technology preservation, technology emulation or data migration. There is a growing wealth of information on digital preservation and related issues available on the Web.

The term ‘accessibility’ as used above deserves further glossing. It is important to distinguish between what we might term ‘machine accessibility’, involving the data integrity of the original files and digital objects, and human accessibility, or the ability to render the object intelligible and usable by a human reader. The latter is the more complex challenge, since it requires not only that all the original bits be present in the right order, but also that there be software available that can interpret them and display them appropriately.

Digitization projects are costly and time consuming, so it is essential that the digitization process should not need to be repeated, incurring more expense and subjecting the originals to further stress. The process may not even be repeatable: some projects we interviewed do not keep the original object. The Library of Congress, for example, converts newspapers to microfilm in certain cases, discarding them when the process is complete. The University of Michigan and Cornell have both discarded original objects for a variety of reasons, such as maximizing shelf space. Now the digital files are the *de facto* preservation copy, so it is even more important that they be suitable for long-term use and remain accessible for future generations.

Preservation should not be considered as an ‘add-on’ to the whole digitization process; the decisions taken throughout — choosing file formats, metadata, and storage — should all consider the long-term sustainability of the digital objects.

Preservation and storage of the digital surrogates

The projects surveyed saw digitization not as a way of creating preservation surrogates from original analog materials, but rather as a reformatting exercise designed to enhance access to and usability of materials. However, as noted above, projects like those at the University of Michigan, Cornell, or the Library of Congress, which discard the analog originals following digitization, must treat the digital version as a preservation surrogate. Furthermore, any project dealing with unstable materials must have similar concerns, since their digital version may in time become the only record of the original. Nonetheless, all of the projects interviewed recognized that they were creating assets that need to be preserved in their own right. Ensuring their long-term availability was particularly important where scholarly work depended upon or referenced them. As the limitations of microfilm become more apparent, many in the cultural heritage community are recognizing digitization as a means of preservation of digital content as well as improving access. Remember that here we are not so much concerned with conservation of the original but with preservation of the digital surrogate itself.

The starting point in planning digital preservation is to identify what can actually be preserved. Various levels of digital preservation are possible, from the ‘bucket of bits’ or bitstore (the storage of files in their original form with no plan to further reformat them for different forms of accessibility or searchability) through to preserving the full searchable functionality of an image or audio system. Each project must decide what is crucial, manageable and affordable. Consider a range of options, from the absolute minimum up to the ideal. If you decide to preserve the minimum, or the bucket of bits, then migration of software and operational systems will be less of a concern. This option might be attractive if your goal is simply to preserve archived data created by others, where your role is merely to keep the data, not to provide an ongoing framework for active use. However, if you decide to preserve the whole system you will need further migration of data to ensure usability in the future. The basic minimum will depend on each project, its rationale and its long-term user requirements. There is no limit to what can be preserved in digital format, apart from the limitations imposed by migration, storage space and other resources, such as staff time.

There are four main issues to consider in formulating your preservation strategy.

The first is software/hardware migration. All products of digital preservation must be migrated at some point, at the very least to a file format that the latest technology can recognize. If you have chosen to preserve the whole system, then operating systems and functional software must be migrated as well. This can cause problems as upward compatibility is notoriously uncertain, even from one version of software to the next, and there is no guarantee that the version you are using will be compatible with releases in many years to come. Full system migration must be carried out frequently to ensure access and usability. You will need to formulate a migration policy that is implemented on a regular basis rather than as a reaction to new software or hardware. Regular testing after migration is also crucial to ensure that functionality has been preserved. This is conventionally

called system emulation.

The second issue concerns the physical deterioration of digital media. All digital media deteriorate over time, but this process will be more rapid if they are stored in an inappropriate way, such as in a damp basement, or as a pile of CDs stacked one on top of another. Correct storage (e.g. in racks that enable the disks to be stored separately) and an environmentally controlled location will help to optimize their stability and protect them from loss. The Genealogical Society of Utah, for example, stores the archival digital master files in an underground vault in Utah as well as in another state in the USA. Tape storage is useful, as there is strong evidence for the extended lifetime of this medium when stored correctly; remember that all tapes must be spooled regularly to avoid sticking. As with software and hardware migration, digital media should be checked and refreshed regularly to ensure that the data are still readable, and this process should be part of your preservation policy. All institutions should draw up a policy that reflects the value of the digital material and therefore sets out how regularly the media should be checked and replaced. This aspect of the preservation policy is linked to hardware migration, as some media formats may no longer be readable by the latest computer hardware. Preserve your data on a medium where the hardware exists to transfer to later media if the original becomes obsolete. Remember that it is costly to use a data recovery agent to move files from an obsolete medium, so make sure your preservation policy will prevent this happening, and migrate while the process is still straightforward.

The third issue concerns metadata, which is crucial to preservation of digital resources. The level of metadata recorded and stored will, once again, depend upon what you choose to preserve, the minimum content or the fully functional system. The format of the metadata is also important, as the metadata should be as accessible as the data. ASCII will not need to be migrated, so it is the most durable format for metadata, although it lacks the functionality that a database might give. SGML/XML files (including HTML) are stored in ASCII format and provide a high level of structure and functionality without requiring proprietary software. The site structure and the relationships between the files should be recorded with a METS format XML file, and the other kinds of files (e.g. images, databases) that make up the site should, if possible, have preservation copies made. Word-processing files should be copied as ASCII text files, and if possible converted to XML to preserve their structure. All databases and Excel spreadsheets should be exported as comma-delimited or tab-delimited ASCII text files. PowerPoint files should be exported to HTML with gifs, with those gifs potentially re-saved as separate TIFFs. GIS datasets should be exported as comma-delimited ASCII.

A fourth and more complex issue is the question of user needs and preferences, which may cause certain formats to become effectively obsolete even while they remain technically functional. For instance, user discontent with microform readers threatens to make microfilm obsolete even though it still fulfills the original goals of its creation. User acceptance—and its decline—will be one of the key “trigger events” that will compel migrations to new delivery versions of digital collections.

Technical, media and file formats

Careful consideration of the technical issues will help you to ensure the long-term accessibility of your digital material.

As outlined above, the physical media on which the digital material is stored need to be managed in monitored, controlled systems and environments to avoid physical deterioration through human mishandling and inappropriate storage. Magnetic media such as digital audio and video tape are highly vulnerable to fluctuations in temperature and humidity, and of course to magnetic fields. Optical media such as CD-ROMs and DVD-ROMs are more durable, but should still be stored with care. For all media, it is advisable where possible to store a second full set in an off-site location, to guard against the risk of theft, fire, and similar disasters.

Choice of medium is equally important. Formats change rapidly and obsolescence is a perennial problem; it may be very difficult in the future to find a drive that can read ZIP disks, just as now it may be difficult to find readers for certain kinds of magnetic tapes. No one format can be guaranteed to persist and remain accessible in years to come, so you need to anticipate continuing developments and new products. Among currently available media, DVD, DLT (Digital Linear Tape) tape and CD-ROM are the most popular. Tape has a tried and tested longevity, while DVD and CD-ROM are newer media and current research has not yet agreed on their likely lifespan. Best practice in preservation is to migrate data from one medium to another, for example from optical disc to tape, while the hardware and software are still available. Migration is an integral part of any digital project and should be set out in your preservation policy.

If projects choose to make back-up or archive copies on fixed media, then it is crucial to ensure that the fixed media is stored correctly and that migration to the latest type (e.g. DVD) is carried out regularly.

A similar obsolescence problem will have to be addressed with the file formats and compression techniques you choose. Do not rely on proprietary file formats and compression techniques, which may not be supported in the future as the companies which produce them merge, go out of business or move on to new products. In the cultural heritage community, the *de facto* standard formats are uncompressed TIFF for images and PDF, ACSII (SGML/XML markup) and RTF for text. Migration to future versions of these formats is likely to be well-supported, since they are so widely used. Digital objects for preservation should not be stored in compressed or encrypted formats.

While TIFF and PDF are *de facto* standards, it must be remembered that they are proprietary formats with multiple 'flavors.' Not all image software can represent TIFF files accurately and Acrobat is required to view PDF files. Projects should monitor the changes in these formats and ensure that users will be able to access the files over time, either by migration or emulation.

Storage and retrieval

The readability of your files is dependent on application software and operating systems. Consideration of file format selection must take this into account. Access software, operating systems and user interfaces are subject to the same process of continuing development and change as storage media and file formats. New versions of operating systems will affect the usability of your files, even though the file format itself may still be accessible. Remember that your operating system may well be proprietary software; unless you are using a non-proprietary system such as Unix, you will have to purchase the more recent version and migrate the files.

It is advisable to separate the underlying content from the discovery and display systems.

Migration will be more complex for those projects that have chosen to preserve the complete integrated system, with full searchable functionality, as there are more features to capture and preserve. Backwards compatibility is often limited and so, as with media migration, these programs should be monitored closely to enable migration before obsolescence makes this impossible. For example, you may be using proprietary software which has regular (often yearly) upgrades. The versions may eventually change so much that backwards compatibility is no longer viable, resulting in, at best, a long and expensive process of data extraction and migration, and at worst, the loss of the data altogether. To ensure accessibility, migrate every file to every new version. The use of standard data formats such as SGML and XML can be an advantage here, since it keeps the data independent from the delivery software. Even if the latter becomes unsupported or if cost forces you to switch to a different delivery system, your data will remain intact and portable.

Link Box:

Preservation of Electronic Scholarly Journals: http://www.diglib.org/preserve/presjour.html

This project, a collaboration between CLIR, CNI and DLF working with the Andrew Mellon Foundation, aims to develop plans for affordable preservation strategies that will implement the OAIS reference model.

Monitor the user interfaces, too. Scripting languages, plug-ins, and applets may all become obsolete over time and users will not be able to access your materials.

In general, remember that standards are the key to creating a resource that is interoperable (the ability of the resource to be used by a different system) as well as sustainable over time. Relying on proprietary software or formats that are not accepted community standards may mean that you will have to migrate more frequently and in isolation. All projects will have to migrate their digital files at times, but by using standards this process will be less frequent and may take place when all projects are migrating data in response to a general change, such as that from SGML to XML, making it easier for everyone.

Organizational questions

The long-term view taken by institutions will also influence the sustainability of digital objects. Projects that have

established preservation policies and procedures, such as Cornell and the Online Archive of California, are identifying preservation issues and placing them at the heart of their institution's ongoing procedures. Insufficient institutional commitment to long-term preservation can create digital resources with limited sustainability.

Similarly, human and financial resources play a huge role in implementing such policies and procedures. Staff turnover and variations in usage will influence the level of commitment to individual projects. Unsatisfactory record keeping and metadata, administrative as well as preservation and access, can also contribute to institutional neglect of digital materials. Moving from individual projects to organizational digital programs will facilitate the long-term preservation of digital materials, by providing greater organizational stability and the overhead necessary to maintain good records and documentation.

Addressing the issues

Whether you have opted to preserve the minimum content or the whole discovery and display system, policies should be in place to ensure the long-term sustainability and accessibility of the digital files you have chosen to be preserved.

We found that attitudes to digital preservation vary enormously. Many of the major digitization projects are using research sources, such as D-Lib Magazine (http://www.dlib.org/), to help them understand the trends and strategies in digital preservation. At the same time, we found that others did not know what digital preservation was and were uncertain of its importance for their projects. There are many questions and no single answer to the preservation issue. No one project is satisfied that it has found the best storage and migration policy to ensure longevity. The Genealogical Society of Utah project is taking a larger view of the issues through its research with Brigham Young University to identify storage media and standards that will help ensure longevity of digital surrogates. All the projects agreed that adopting standards in creating digital objects would enable their sustainability. Each project will choose what it will preserve based on the level of resources available for storage and migration as well as user needs.

Research is being carried out that will help you make the right decisions for your digital resources. Projects, such as CEDARS, have attempted to establish strategies for ensuring digital preservation through technology preservation, technology emulation and data migration. VidiPax and other companies are addressing the technical media side through research into restoring and archiving digital media, while organizations such as NARA are addressing both traditional concerns (records management) and new technical standards. Internationally, The Digital Preservation Coalition in the UK is working to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally http://www.jisc.ac.uk/dner/preservation/prescoalition.html, and Preserving Access to Digital Information (PADI), a project of the National Library of Australia, maintains a useful web site with international research results and recommendations (see Link Box below).

The key to understanding and implementing digital preservation is to view it as a series of tasks rather than as one process. These tasks, if properly carried out, will help ensure not only that digital objects are correctly stored but also adequately maintained and useable over time.

Although there are others, six methods of digital preservation are widely discussed among heritage professionals. They are not mutually exclusive, and you may decide on one or a mixture of methods.

Technology preservation relies on preserving the technical environment in which the digital objects were created, both hardware and software. This approach is more relevant to projects that have chosen to preserve the full system with functionality and interface as well as the actual content. The digital files must be preserved as well as the computer and software. Maintaining hardware for long-term use is problematic, and it is also costly to store the physical machinery as well as the digital files. This method would apply to projects that wish to maintain the environment for historical use (a computer museum, for example) and may not be appropriate for many cultural heritage projects.

Technology emulation also seeks to preserve the technical environment, but instead emulates the environment on current technology, mimicking the software and even the hardware. Everything has to be emulated, including operating systems, scripts, and applications. This is unlikely to be a viable approach for most digital imaging projects. HATII's Digital Archaeology Report for the British Library (http://www.hatii.arts.gla.ac.uk/Projects/BrLibrary/index.html) has more detailed explanations of emulation.

Data migration focuses on maintaining the objects in current format and may involve migrating all files to a

newer format (for example in 2001, from JPEG to JPEG 2000), when current software no longer supports the earlier file format. Data migration is a viable option for cultural heritage programs and will involve making certain decisions, such as how often to migrate and what format to migrate to. It can be a time-consuming process, involving inevitable loss of data as well as sacrificing an element of the 'look and feel' of the original material. Using community standards at least gives the reassurance that every project will be migrating data of similar types at roughly the same point in time, making it more likely that tools and support will be available. The Archivo de Indias project has demonstrated that it is possible to migrate millions of documents form one file format to another, provided you design the work program effectively.

Enduring care encompasses all aspects of caring for the digital objects and the media they are stored on, including housing in secure locations and careful handling. This is short-term care and projects must consider migration and refreshing for longer-term preservation.

Refreshing is the process of moving data from one medium to another, e.g. from CD-R to DVD. It is not a total guard against obsolescence and should be seen as part of the whole digital preservation strategy.

'Digital Archaeology' describes what to do in a 'worst case' scenario, where a digital preservation policy has not been followed, or an unforeseen catastrophe has damaged the media. As described in the HATII report on Digital Archaeology, this is the process of rescuing content from damaged media or from obsolete or damaged hardware/software.

Storage Media

Data migration should not be confused with *'data refreshing'*. Data migration is the act of moving the actual content of the digital files to a newer file format (e.g. from Access 97 to Access 2000), while data refreshing is moving the files (no matter what format they are in) to a new physical medium, e.g. from CD to DLT tape, either because of obsolescence (e.g. laser discs) or because of physical deterioration.

Each project will have to identify the best storage strategy for its digital masters. This will include the choice of medium and the choice of institution/organization entrusted with managing that medium. The benefit of choosing a management service is that you can forego the responsibility and effort in choosing the medium. The benefit in investing the time to choose the medium is that you enjoy total control over your collection. Favored media vary from project to project and depend upon costs, access issues, physical storage space and technical equipment available. An in-depth examination of the technical issues of digital media is beyond the scope of this *Guide*. For a fuller explanation of these issues see the HATII report on Digital Archaeology cited above.

The formats used by the projects interviewed for this *Guide* include:

CD-ROM (including CD-R)

DAT Tape

DVD - both DVD - Video and DVD- ROM

DLT Tape

RAID Server

A mixture of the above media

Projects using audio-visual material will need to use a medium that can deal with the large file sizes created when digitizing such materials as well as enable access to the stream of data. DVD and DV Stream are the most popular media for these formats.

CD-ROM is a popular storage medium — it is cheap and simple to process. Projects such as the Online Archive of California, the Digital Imaging and Media Technology Initiative of the University of Illinois and the Colorado Digitization Project use CD to archive their digital materials. CD writers are easy either to install and run from the machine that is temporarily storing the data, or to back up from the server that holds the access data. CD-R has less storage capacity than DLT tapes or DVD — CD-R can hold approximately 640 MB of data while a DLT tape holds 35-70GB. Magnetic tape is known to be stable for approximately 30 years under good storage conditions. DLT tapes for instances should survive for a million head-passes over a 30 year period. Manufacturers claim that CD-Rs will last from 50 to 200 years, but experience indicates that even after as little as 10 years they are beginning to suffer from drop-outs. Media lifespans are an unknown quantity, so you may count on having to

migrate and refresh the data regularly. These figures are based on climate- and temperature-controlled storage conditions and safe environments. Lifespan also assumes that the media are free from manufacturing errors and are not mishandled.

You can act to improve the storage life of your media. We recommend that you handle media with care and that you check it regularly. The following two tables give examples of how to improve the lifespan of various storage media.

Good Practice Box:

Improving the lifespan of CDs and DVDs:

Avoid	Never	Always
Damage to the upper and lower surfaces and edges of the disc	Attach or fix anything to the surface of the CDs	Store media in a jewel case or protective sleeve when not in use
Scratching and contact with surfaces that might result in grease deposits (e.g. human hands)	Write on any part of the disk other than the plastic area of the spindle	If using sleeves, use those that are of low-lint and acid-free archival quality
Exposing discs to direct sunlight		Wear gloves when handling the master CDs

Good Practice Box:

Improving the lifespan of DLTs:

Avoid	Never	Always
Placing the tapes near magnetic fields	Stack the tapes horizontally	Keep tape in its protective case when not in use
Moving the tapes about	Put adhesive labels on the top, side or bottom of cartridge	Move tapes in their cases
	Touch the surface of the tape	Store the tapes in appropriate
	Put a tape that has been dropped in a drive without first visually inspecting it to make certain that the tape has not been dislodged or moved	Store the tapes vertically

There are two types of DVD: DVD-Video and DVD-ROM, which can be compared respectively to the audio CD and the CD-ROM. DVD-Video can contain a whole film with digital audio. The DVD-ROM stores computer-readable data.

Many larger projects store two copies of the media, in different storage facilities. Both the GSU and SHOAH use

underground vaults with temperature controls and high security in two geographically remote locations. SHOAH has the added risk factor of earthquakes and has chosen its second site in Washington DC. Not all projects can afford such facilities and the additional security that they provide, and will simply store one copy on site. We recommend, however, that all projects store archived master files off-site where possible to ensure against natural and human disasters.

We found that many projects did not check archived data for refreshing and did not have a preservation strategy at all, although many are developing them.

Stargazing Box:

There is a need for better long-term storage formats. The GSU project is conducting very interesting research into media for digital preservation. Through their work with BYU and Norsam (http://www.norsam.com/), they are examining media such as the HD-ROM, which has the following manufacturer's claims:

HD-ROM and the "ET" System are planned to have the following specification features: (ET is Electrical Transport Data Storage System)

- 200 Gigabytes per CD-sized disc
- 2 terabyte 10-disc Modular Reader
- 60 terabyte Modular Storage Unit which holds 30 10-disc Modular Readers
- Write rate of 30 megabytes per second
- Expandable to petabyte storage
- High-speed random access
- Each disc will have its own read head
- Read rate comparable to CDs

File formats

The file formats you choose are crucial to ensuring data migration. In 2001, uncompressed TIFF is the preferred raster image format for digital masters. Migration from this format may be inevitable to ensure long-term access to the materials. However, TIFF is widely supported and software manufacturers may continue to support it for the foreseeable future. It should be monitored carefully. TIFF is platform-independent. Archival copies should be stored uncompressed.

JPEG, JPEG 2000, MrSid, and GIF formats—which may be well suited for delivery—are not considered to be as well suited as TIFF for preserving digital master images, for two main reasons. All of these formats incorporate compression algorithms that offer potential benefits for storage and networked delivery, but increase the requirements for both monitoring and migration. Compression is one more thing to worry about when tracking the availability of software that can read and write formats without introducing information loss; thus, it is possible that intervals of migration will be more frequent for compressed formats rather than for uncompressed TIFF. In addition, because these formats employ lossy compression algorithms, there are opportunities for information loss at the time of capture and in subsequent migration activities. Other formats with lossless compression, such as PNG, are not widely used yet because of their relatively limited support within software packages and image viewers. TIFF, while not an open standard in the strict sense, is the *de facto* standard format for digital masters within the community, but not necessarily for delivery, given the fact that standard web browsers will not display TIFF images.

Formats for audio-visual material include WAV, AIFF, MPEG (various versions) as preservation archival copies. RealMedia (both audio and visual) and QuickTime are used for streaming data to users but not as archival formats.

Encoding for longevity

The structure of digital objects is just as important as media and format in ensuring their longevity.

Digital objects stored in a proprietary format, such as a proprietary database, are difficult to migrate forward;

proprietary software often has a short lifespan and is vulnerable to changes in the company that owns and supports it. A proprietary database can be unreadable within five years. You can minimize this risk by exporting your data in a standard format (such as tab-delimited ASCII). Similarly, use standard encoding schemes to structure your digital objects in order to minimize the risks of data obsolescence. Encoding standards such as SGML or XML are platform- and software-independent, and many projects, including some of the larger ones such as those at the Library of Congress, Cornell, Michigan, the Online Archive of California, and Virginia, have adopted these formats.

Metadata

One of the most important factors in ensuring digital preservation is the creation of robust metadata to describe the process of digitization as well as the actual digital objects themselves. Administrative metadata are particularly associated with digital preservation and the retrieval of the images from the archived media. This facilitates both short-term and long-term management and processing of digital collections and includes technical data on creation and quality control. Administrative metadata also handle issues such as rights management, access control and user requirements as well as preservation strategy information. The administrative metadata should include technical data that can be of potential use to systems (e.g., for machine processing tasks during migration), and to people, such as administrators monitoring the number and types of formats in the repository, or even to users inquiring about the provenance of a digital object. Having metadata in place will facilitate digital asset management, by helping to manage the digital files, their location and their content. Also see Section XIII on Digital Asset Management.

Link Box:

Some projects are adopting an institutional approach to metadata.

- Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/
- CEDARS project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
- Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metatada: http://www.nla.gov.au/padi/topics/32.html
- NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
- OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
- Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf
- Amy Friedlander, “The National Digital Information Infrastructure Preservation Program: Expectations, Realities, Choices and Progress to Date,” D-Lib Magazine (April 2002) Volume 8 Number 4; http://www.dlib.org/dlib/april02/friedlander/04friedlander.html
- OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
- A Recommendation for Preservation Description Information: A Report by The OCLC/RLG Working Group on Preservation Metadata: http://www.oclc.org/research/pmwg/pres_desc_info.pdf
- Anne Kenney et al., “Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism,” D-Lib Magazine (January 2002), Volume 8 Number 1; http://www.dlib.org/dlib/january02/kenney/01kenney.html

While administrative metadata are crucial to the long-term usability of digital materials, structural and descriptive metadata are also critical to the digital preservation process. All projects recognize the importance of good, robust metadata and have implemented schemes that facilitate the maintenance and accessibility of the resources. The importance of high quality metadata cannot be underestimated, as this can ensure that while data migration occurs, the original information about the digital object is not lost and can be recreated if necessary.

Organizational strategies

Over the last few years, those involved in developing and delivering digital objects have realized that preservation issues go far beyond questions of preserving bits and dealing with obsolete media and file formats. Indeed, there is widespread recognition of the need to develop trusted digital *repositories* that can ensure not only migration of media and formats, but also the preservation of relevant information about digital objects so that we can preserve their behaviors as they are migrated forward in time and make meaningful use of them in the future. We need to preserve not just the byte stream, but also the structure of the digital object and its context. Discussions about the attributes of digital repositories have crystallized around the emerging ISO standard, "Reference Model for an Open Archival Information System," originally developed by NASA to deal with the huge volume of digital material developed by the space program. (In May of 2002, RLG-OCLC issued a report based upon this model entitled "Trusted Digital Repositories: Attributes and Responsibilities.") The Draft standard states:

> An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.... The model provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives.

OAIS is a reference model, not an implementation plan. Libraries and archives are just now beginning to design digital repositories. [1] As they look to implement OAIS systems, metadata concerns loom large. Long-term preservation of digital objects in a repository requires that we develop and maintain structural and administrative as well as descriptive metadata about those objects. The METS initiative has gained much attention this past year as a flexible way to address these needs. The FEDORA project, which is complementary to METS, is aiming to develop behavior metadata as well. At this time, creators of digital objects and projects do not yet have full guidance on what digital repositories will need to secure their objects for the long term. In particular, much of the progress to date has focused on ingesting, storing, and preserving individual digital objects and much more research needs to be done on the long-term needs of complex digital creations. Nonetheless, attention to the emerging structure of repositories and their metadata needs will definitely increase the likelihood and ease with which digital objects can be incorporated into repositories and preserved for long-term access.

One of the major issues involved in creating trusted repositories is that of responsibility. To what extent will creators and local providers need to cede control over their digital objects to centralized repositories? The advantages that large, financially secure institutions can provide in terms of long-term preservation are unquestionable. Nonetheless, research efforts especially amongst the digital library community are also focused on developing distributive systems. Indeed, there is a growing sense that "lots of copies" will be an important preservation and authentication strategy for the foreseeable future. With the ever-declining cost of storage space, redundant distributive systems with shared responsibilities has many advantages that mitigate against centralized control.

Conclusion

Digital preservation is a discipline that is still in its infancy. Much of the preceding discussion provides a general description of various conceptual approaches to digital preservation, and also provides guidance for good custodianship (in the present) of existing digital files and digital media. However, it is expected that, in the near future, digital preservation will be executed within the context of enterprise-wide digital repository systems. This is an area of ongoing research and development. For example, in the United States, there is the new National Digital Information Infrastructure Preservation Program (Friedlander 2002), whereby the Library of Congress will lead a national planning effort for the long-term preservation of digital content and will work collaboratively with representatives of other federal, research, library, and business organizations.

Summary

Key factors in preservation are:

- Identifying the data to be preserved — basic content and metadata through to the fully functional program with interface look and feel
- Adopting standards for file formats
- Adopting standards for storage media
- Storing data on and off site in environmentally secure locations
- Migrating data
- Refreshing data
- Putting organizational policy and procedures in place

---

[1] Examples of projects using OAIS are CEDARS, NARA, NEDLIB and PANDORA. A chart for the deposit system for electronic publications can be found at: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html. The CEDARS project aims to produce strategic frameworks for digital collection management policies, and to promote methods appropriate for long-term preservation of different classes of digital resources, including the creation of appropriate metadata.

table of contents previous chapter

valid xhtml 1.1
abp~03/03

VI. Capture and Management of Images

Introduction

Images have tremendous power on many levels and in many contexts. Before the advent of the Web, the online world was largely limited to textual data, but with the growth of the Internet, the demand for images is huge and continues to grow. In addition to the compelling visual pleasure they offer—which puts them in high demand in the culture generally—they also carry dense cultural meaning and are being used increasingly for research, teaching, and learning.

As a creator and distributor of digital images you have many factors to think through.

As with other kinds of materials, these include the selection of objects and visual documentation for digitization, the intended purpose of the digitized items or collection(s) and their expected users and uses, as well as the availability of relevant resources. Your approach will be determined by a number of considerations: the nature of your material, the needs of your users, and your available financial resources and human expertise. Overall, you should remember that your objective is to create a high quality digital resource that will remain useful in the long term. This resource should have as its cornerstone an archived digital master that can be considered "use-neutral" and can serve as the parent to many children that are used in print, web, video and other forms of reproduction.

Other sections in this Guide provide information that bears directly on the issues in this section. Section II provides a broad contextual view on overall project planning, including many of the larger decisions that affect image capture. Section III gives guidance on selecting materials, which will play a crucial role in determining what kinds of image capture techniques are appropriate. Section VIII offers detailed information on quality control and assessment, and Section XII discusses user needs evaluation, which can help you decide issues such as the required image quality. With this information as a backdrop, an overview of the steps to be taken will include the following:

- Investigate the formats of the source material (works on paper, original paintings, slides, negatives, transparencies, photoprints, maps, postcards, etc., as this will affect your choices of method and equipment and your need for specialized expertise. It is likely that the project contains multiple types of material that need to be evaluated both separately and as a group when planning the scope of the project, the equipment needed, and the most effective work flow.
- Establish an understanding of the characteristics and properties of your source material (its fragility, size, mobility, access, etc.) as this will also affect method, equipment, costs and expertise.
- Assess the way it conveys information (how important is color, line, detail, text?) and the kinds of metadata that can or must accompany it.
- Decide whether you will digitize in black-and-white or color (a variety of bit-depth decisions need to be made).
- Decide on whether you have the expertise needed in-house, whether you need some training or whether you will outsource the job.
- Decide on file format, size, storage and management.
- Define benchmarks to ensure image capture consistency across the collection.
- Test a sample of images for capture, display, and print output.
- Decide on standards for documentation (metadata) of the digital files.
- Clarify decisions about sustainability of the project and preservation of the digital files.

Finally, we emphasize that this chapter should be read alongside the chapter on Quality Control and Assurance, which is an integral element of the creation process.

Planning for Image Capture

Choice of Source: Original versus Intermediate

During the selection stage, you will have chosen the items to be digitized on the basis of their content or information value as well as for their potential importance to a set of users. At this stage in the planning you should also consider whether to digitize from the original object or from an existing surrogate, for instance if the original is inaccessible or could suffer damage from handling during the process. Examples of derivative formats are photographic prints of original art works, microfilm of newspapers, and slide images of objects.

Frequently, the physical size and condition of the object will determine this question. If an original artwork is oversized, then a flatbed scanner will not work. In addition, some objects should not be placed in contact with glass. For example, fragile work, determined as such by a conservator or preservation specialist, should not be put in contact with a scanner platen or a glass plate holding materials flat on a camera copystand. It may be possible to digitize such objects if a high-end digital camera is in place, with the requisite equipment, such as book cradle and cool lights. Even “cool lights” may emit harmful ultra-violent radiation or destabilize the temperature and humidity in the room, and if this will put the objects at risk you may need to look into the latest generation of electronic flash-based, area array digital cameras. If such equipment is not available, digitizing from a film intermediary may be a possible alternative. With any intermediary there will always be some information loss, so digitizing from the original is advisable when the highest level of quality, per item, is needed, in cases such as preservation replacement. It should be noted that digitizing from the original can significantly raise costs—not only for capture, but also for preservation—which is why intermediaries are used so widely in large-scale, production-level digitization projects.

Intermediaries pose their own problems, including poor production and scratched, faded, or out-of-focus images, all of which affect the quality of the digital product and need to be closely monitored through the quality control and assurance process. However, staff at some projects included in our survey, for example those at the University of Michigan, photograph all the originals and digitize the resulting 4” x 5” transparencies. This approach is not common among the major projects, and most will digitize from the originals. It should be noted, though, that it is better to do a good job of digitizing a well-made intermediary than a merely adequate job of digitizing original images.

Before deciding to use intermediaries, it is worth considering the following factors:

- Condition of the object: is it rare and fragile? will it be damaged by exposure to intense or unfiltered light?
- Physical size of the object: is it too large to be managed on the available equipment?

If the answer to these questions is yes, then you should consider using a film intermediary.

Quality and Resolution

Decisions as to what resolution, bit-depth or method of tone reproduction, and image format to use for capture, storage and delivery can only be made after project staff have completed an assessment of the material, the needs of potential users, the available technology, costs per image and likely financial support.

The key questions will be:

- What information content do you wish to convey and in what format(s) – screen image, print, or both?
- What functionality does the end product need to support? For instance, does it need to accommodate art history research, which might require scrutiny of minute details such as brush strokes at high magnification? Or is it intended for less exacting audiences where download speed is more important?
- What capture standards and equipment will achieve this?

While you could produce a matrix and try to balance all these parameters, experience shows that you will inevitably need to make trade-offs. Although local circumstances such as resource constraints, long-term plans for the material, and other factors will dictate what trade-offs are appropriate, other things being equal it is probably a good rule of thumb to reduce quantity and maintain quality. A higher-quality digital object offers greater flexibility for future use, since you can always derive lower-quality versions from it, whereas if your needs become more demanding, a lower-quality image will be inadequate. This consideration may carry additional weight if you are dealing with rare or perishable materials, where you may not have a second opportunity to digitize in the future.

There is no set resolution for any individual resource or collection. Each project must identify the minimum level

of quality and information density it requires for its digital surrogates. A detailed explanation of sampling rates and other measures of information density can be found in the appendix on digital data capture. Briefly, resolution is usually expressed in dots or pixels per inch (dpi or ppi) and measures the density of sample—the number of information samples taken per unit of area— that is captured by the scanning equipment. Generally speaking, the higher the ppi, the more detail being captured. The sampling depth, or bit-depth, measures how much information is captured in each sample: for instance, how many colors or levels of grayscale. The higher the sampling depth, the more subtle the gradations across the image, and the larger the resulting file size. The sites interviewed for this Guide used a range of sample rates from 300 dots per inch (dpi) to 600 dpi for their master image files. Lower resolution images were often derived from the master file to produce distribution copies with smaller file sizes for ease of download. The most widely adopted format for storing preservation quality digital masters is uncompressed TIFF. Derivatives, in such file formats as JPEG, are used for access and delivery versions of the image. The resolution of these derivatives generally ranges between 72 ppi and 150 ppi. The resolution appropriate for a high-quality image of the source material should be determined by the size of the smallest significant detail that will need to be visible, given the intended use of the image.

Images can be either reflective (e.g. photographs or drawings) or transmissive (e.g. transparencies or negatives). With reflective material, light is bounced off the surface of the image and its intensity measured using light sensitive diodes. In the case of transmissive material, light is instead passed through the original source. There are three constraints on the highest resolution you can use in scanning both reflective and transmissive material: the maximum sampling capabilities of the scanning device; the size of the files you can manage; and the level of information you actually need. An example of this last point involves old postcards, which were often printed on poor quality paper using cheap processes. If scanned at high resolutions you capture the texture of the paper, which can obscure other information in the image.

Remember that before digitizing you still need to answer questions about the format of the material, its size (or size range if in a batch); whether you need color or black and white (see the discussion of this question in the cost-benefit analysis section of Section III: Selecting Materials); what is the smallest significant detail in the image that you wish to capture (fine line drawing, detailed photographs, fine art); and to what range of uses will the digital objects be put?

Benchmarking

The greater the detail that needs to be captured, the higher the resolution required. Many projects use benchmarking to identify the best resolution for capturing a given image. Benchmarking will identify the smallest essential detail by investigating the attributes of the source and will address the issues of whether the objective is to convey the informational content of the original or to convey information about its creation, such as brush strokes or engraving methods.

One way to establish a benchmark is to take a sample of the material, perhaps that with the greatest detail, and scan at a variety of resolutions. Show these scans, using the medium on which they will be chiefly viewed in practice, to a sample of users and staff to identify which sampling rate best meets the needs for the digital object. This process will be most effective if you take into account the kinds of equipment your users will actually be using: will they have access to high-end monitors and fast broadband connectivity? Once you have identified the optimum resolution then you can set the specifications for that collection. An alternative to 'evaluative benchmarking' is to use a Quality Index (QI). The Cornell Digital Quality Index (DQI, http://www.library.cornell.edu/preservation/tutorial/conversion/conversion-04.html), developed from guidelines for the micrographics industry, though currently restricted to "printed text, line art and book illustrations," can be used for bitonal and grayscale scans, by measuring the smallest meaningful area of the image and using a QI to calculate the optimum dpi.

The UK's Higher Education Digitization Service (HEDS) has a Matrix for Potential Cost Factors (http://heds.herts.ac.uk/resources/matrix.html) which can be useful when considering different resolutions. Another good resource to refer to is the California State Library Scanning Standards (http://www.library.ca.gov/assets/acrobat/scandocrev1122.PDF) but you should use such instruments with great care.

The main principle to remember is that, once the resolution has been identified for a particular batch or collection, it should be fixed for those materials for the duration of the project.

Method of tone reproduction

When you undertake the actual scanning of your images, you will need to choose between several methods of tone reproduction. This phrase refers to the number of bits (or “bit depth”) sampled for each pixel, and to the technique used to capture the full range of tones (or density) in the source materials.

Scanners record tonal values in digital images in one of three general ways: black and white, grayscale, and color. In black-and-white image capture, each pixel in the digital image is represented as either black or white (on or off). You can choose the threshold at which a given point will be considered to be black, and you can use halftoning algorithms to create a “screened” pictorial image, but black and white scanning is generally appropriate only for digitizing text and line art. In 8-bit grayscale capture, the tonal values in the original are recorded with a much larger palette that includes not only black and white, but also 254 intermediate shades of gray. Again, the thresholds can be controlled—either manually or automatically—as a first step to ensure that meaningful information in the highlight or shadow areas of images are adequately captured, and to get optimal results for originals with different degrees of contrast and overall darkness. In 24-bit color scanning, the tonal values in the original are reproduced from combinations of red, green, and blue (RGB) with palettes representing up to 16.7 million colors.

Your decision about which method to use for tone reproduction begins with the choice of black-and-white, grayscale, or color scanning as the default minimum for various classes of material. Your user needs will be an important determinant, but you should not assume that there is a simple correlation between increasing information capture and increasing user benefit. If your original is a faded black-and-white photographic print, for example, color scanning might accurately reproduce the effects of aging. However, file sizes for the color images would be three times larger than their grayscale counterparts. If color is not needed to meet the stated viewing, printing, and publication objectives for your project, grayscale may be the better option since it will be faster to download and smaller to store. Relatively speaking, adjusting resolution is a very straightforward process. Adjusting the settings for tone reproduction, however, can become complicated. Assume that each incremental jump might require changes to equipment, to operator skill, or both.

Having selected bit depth, the most important decisions regarding tone reproduction relate to reproducing the densities of tone in the source materials. Good digital masters are characterized by having pixels distributed across the tonal range in the image from black to white. Stated in the negative, tones are not “clipped” in good masters. Following scanning, all subsequent image processing relies upon pixels either being in the region of interest or not. A scanner that automatically chooses to reproduce the middle range of tones in an image might not have enough dynamic range to capture the meaningful highlights and shadows at both sides of the spectrum. Because experts say that much of the secret to good color reproduction is getting the neutrals right, the method used to capture densities should be carefully planned. For reflective materials such as photographic prints, grayscales might be used to set the “aimpoints” for black, white and middle gray in the digital masters. With transmissive materials, however, these decisions might need to be made by the operator on a case-by-case basis, which is one of the reasons that it can take much longer to scan negatives than prints in order to produce pleasing reproductions.

In addition to user needs, you should also consider which method will actually capture the information of your originals best. Grayscale imaging is appropriate for continuous tone images which lack color. Black-and-white imaging might be warranted for line art if cost savings are a concern; for instance, a black-and-white scan of a typescript page might provide adequate legibility. However, many forms of line art (for instance, manuscript pages) contain subtle color variation which might be of importance in distinguishing overlays of pen strokes and identifying the order in which the marks were made. For these purposes, full color scanning would offer significant benefits.

Projects will occasionally benchmark their imaging by establishing a minimum resolution for the collection as a whole. Most often the resolution depends upon the original, the resources available and the use that you intend to make of the material. Sample testing provides a good way to benchmark sampling rate. Setting minimum requirements is generally accepted as the best practice for establishing the capture requirements, but this must be weighed against the resources available.

Projects with very long-term digitization efforts may have to address the changes in standards for creating digital resources. Improvements in scanning technology and the diminishing cost of storage have radically increased the cultural heritage community’s minimum standards for capture quality, and digital objects captured ten years ago may have been scanned at what would now be an almost unacceptably low level. A museum or library with a large number of legacy images captured at lower resolution will need to decide how to handle the transition. One possible strategy would be to digitize new material at the current recommended standard, and create from these new scans a set of lower-resolution derivatives to use with the legacy materials, if internal consistency is a crucial consideration. Over time, it may be possible to rescan the legacy materials and slowly upgrade the entire

collection. Under most circumstances, it would be a waste of time to scan the new materials at an arbitrarily low resolution. The conversion of material from analog to digital will always be done in an environment of advancing technology and improving standards. You should be sure that you strike a balance between the costs, preservation suitability of the format, and the best usable standards. Generally, best practice is to capture at as high a level of overall quality as is appropriate to handling, quality, budget, or use considerations in order to limit the upgrading that might be required otherwise.

The advice box below gives summary guidelines for selecting the mode of tone reproduction.

Practical Advice Box:

Some practical guidelines for selecting the appropriate scanning method

- Use black-and-white (1-bit) scans if the image is black and white and contains no shading.
- Use grayscale (8-bit) for material with shades of gray or continuous tones: black-and-white photographs, half-tone illustrations, typescript, archival materials in black and white where, say, ink density varies and paper tonality appears to be important.
- Use 24-bit color or higher whenever color is meaningfully present in the original.[1]

Some projects scan at higher than 24 bit-depths (32-bit color 4 color channels of 8 bits each captured as red, green, blue, and 8 bits of grayscale or 48-bit color) even though the current generation of applications cannot render this depth at the moment. This is a hedge against future improvements in technology that you might wish to adopt. For originals of wide color gamut—e.g. paintings or natural history specimens—it may be important to select capture devices that allow the storage of raw files of 12, 14 or 16 bits per pixel, if an extremely high degree of color accuracy is a goal. Dynamic range may be checked and the image archived for future output devices that will be supported. Document the specific characteristics of the capture device, e.g.. non-interpolated highest resolution, spectral response captured IT-8 targets with the 50% gray neutralized, etc.

Transmissive material (for instance, slides and negatives) requires special treatment when scanning. The decisions made for resolution and bit depth are the same as for reflective material, but as slides and negatives tend to be smaller in size the resolution must be higher to achieve a digital image of useful size and quality. Although film scanners have generally been recommended to capture transmissive material, today many mid- to high-end flatbed scanners outperform production slide scanners. Projects at the Library of Congress capture 35mm slides up to 3500 dpi (the range is from 1200 dpi to 3500 dpi). Individual projects will identify the resolution required depending on the original source, the detail of content and the users' requirements.

You also need to ensure that there is a match between the dynamic range of the material that you are scanning and the abilities of your equipment to capture that range. Dynamic range is expressed as a scale from 0.0 (perfect white) to 4.0 (perfect black). Some color scanners miss subtle differences between dark and light colors and as a result do not create faithful digital representations of the content. The better quality the scanner, the greater the dynamic range. Drum scanners can capture between 3.0 and 3.8 and give good color quality; however they are very expensive and impractical for large projects. The scanner's ability to capture dynamic range is an important factor when deciding which equipment to purchase and this is especially true if you are working with transmissive material. Do not believe the manufacturer's claims of dynamic range. Instead, make your own tests by scanning a standard 22 step Kodak film gray scale and subtracting the highest from lowest perceptible densities. A scanner might have a high dynamic range but also introduce unacceptable noise in the dark areas of the image. If you cannot do your own testing, you may be able to find another institution that has recorded results and is willing to share them.

Knowing your sources will allow you to ensure that you have configured your equipment appropriately. For instance, halftone images are made of regular patterns of dots or lines (the word "halftone" is derived from the printing technology method of representing photographs or artwork). They can be either color or grayscale. During digitization, the scanner dot sample combined with the halftone dots can produce undesirable patterns known as the 'moiré effect'. One way of avoiding this is to scan at high resolution, but you may also need image enhancement processing, either post process or at scan time. Most good quality capture software packages will enable descreening to minimize this kind of scanning artifacts, although you may have to do this post-processing, using an image processing tool such as PhotoShop.

Process and Management Issues

Choosing a file format

In choosing the appropriate scanning method you must not only identify the appropriate digitization standard but also the best file format. You should select file formats that are platform-independent and supported by a wide variety of applications, both for browsing and editing. It is generally understood that for master images, uncompressed TIFF (Tagged Image File Format) provides the most suitable preservation format. 'Uncompressed' means that all the information encoded during the scanning process is retained in full. Digital projects store their master archived files as TIFF and then create smaller or derivative files for display and access for the users. These derivatives may be created using a number of different formats.

The file formats typically used for display and access are JPEG (which will gradually be displaced by JPEG2000), GIF, MrSid, PNG, and PDF. These file formats use compression algorithms to reduce the size of the digital files. Some algorithms, such as the LZW (Lempel-Zif-Welch) encoding scheme used by GIF do this to some degree without disposing of information altogether. LZW, for example, uses a lossless algorithm to compress files by up to 33% without throwing away data irrevocably. When the file is decompressed it is possible to reconstruct the compressed data. Other approaches to compression, including fractal and wavelet compression, take more radical approaches and sacrifice data to minimize file sizes. JPEG (Joint Photographics Experts Group) takes advantage of 'lossy' compression algorithms, and in particular fractal compression.MrSid (Multi-Resolution Seamless Image Database) uses wavelet-based image compression, which is especially well-suited for the distribution of very large images. The Library of Congress uses MrSid to deliver maps from its collections. As well as having impressive compression capabilities with limited visible information loss, it stores multiple resolutions of images in a single file and allows viewers to select the resolution (in pixels) that they feel will be most appropriate. There are other file formats that are suitable for delivery of images in different environments and you should consider which of these will be most appropriate for your project. The table below provides an indication of the most commonly used formats.

Definition Box:

There is a range of non-proprietary and proprietary image file formats available. This table includes some of the more common formats for raster images.

Extension	Meaning	Description	Strengths/weaknesses
.tiff, .tif	TIFF (Tagged Image File Format)	Uncompressed file. Originally developed for desktop publishing. 1 to 64 bit depth. Used mostly for high quality imaging and archival storage.	Generally non-compressed, high quality. Large file sizes. Most TIFF readers only read a maximum of 24-bit color. Delivery over web is hampered by file sizes. Although LZW compression can reduce these file sizes by 33% it should not be used for archival masters.
.gif	GIF (Graphics Interchange Format)	This 8-bit file format has support for LZW compression, interlacing and transparency.	Lossless compression. Popular delivery format on web. .png was defined to replace GIF.
.jpg, .jpeg	JPEG (Joint Photographic Experts Group)	Compressed images. 8-24 bit. Variable amount of compression to vary quality and file size.	Lossy compression. Widely used delivery format. Flexible.

MrSid	Multiresolution Seamless Image Database	image-compression technology	Lossy compression. can compress pictures at higher ratios than JPEG; stores multiple resolutions of images in a single file and allows the viewer to select the resolution.
.pcd	ImagePac, PhotoCD	Lossy compression. 24 bit depth. Has 5 layered image resolutions.	Used mainly for delivery of high quality images on CD.
.png	PNG (Portable Network Graphics)	Lossless compression. 24 bit. Replaced GIF due to copyright issues on the LZW compression. Supports interlacing, transparency, gamma.	Some programs cannot read it.
.pdf	PDF (Portable Document Format)	4-64 bit depth. Uncompressed. Used mainly to image documents for delivery.	Need plug-in or adobe application to view.
.pct	PICT	Compressed. Mac standard. Up to 32 bit. (CMYK not used at 32 bit.)	Supported by Macs and a highly limited number of PC applications.

File sizes

You will need to determine how much storage space your image files will require and this is best done by working out the file sizes first. There are different approaches for scanners and linear scanning digital cameras. In the Calculating Box, Estimating Approximate File Sizes, we provide you with an example of the method. The quality of the master images will reflect the work processes you put in place, the equipment you use, and the resolution (dpi) and bit-depth (BD) you select. External variables that may limit this are the amount of storage available and the time for scanning each document.

Calculation Box:

Estimating Approximate File Sizes (in bytes):

Approximate file sizes for material created with a flat-bed scanner can be determined using the following formula:

$FS = (SH \times SW \times BD \times dpi^2)/8$

FS = file size

SH = Source Height (inches)

SW = Source Width (inches)

BD = bit depth

dpi = dots per inch

/8 because 8 bits = 1 byte, the unit in which file sizes are measured.

Large file sizes make compression essential for delivery to users. Derivative files in JPEG format are interoperable within the community and accessible through common access systems, i.e. the Internet. In some cases a much smaller thumbnail version of the image is used for quicker access (for instance, in displaying search results), allowing the user to download a larger file once they have found the image for which they are looking.

Calibration

Calibration is a crucial aspect of the quality control process in any digitization project. You will need to calibrate the work environment, the monitor, the capture devices, and the light sources. Frequently during the life of the project you will find it necessary to recalibrate this equipment as well.

First, establish the physical environment where the digitization work will be carried out. Avoid fluorescent lighting, natural lighting, reflections, and bright colors on the walls. Ideally, darken the space with black out curtains or dark neutral walls, to eliminate ambient lighting, so that the only light source is the scanner bulb or the cold lights used for the digital camera. Where possible, maintain a dust-free environment with operators wearing neutral colored clothing and protective shoe wear. Avoiding carpets on the floor reduces dust accumulation. While it may not be possible to meet all these conditions, it is crucial to ensure the absence of ambient lighting — do not set the equipment up in a sunny room with fluorescent lighting! Consult conservators to insure that the temperature and humidity conditions will be stable and appropriate for the original work.

Once you have established the physical space, calibrate the monitor of the workstation at the beginning of every day, to ensure that the contrast, brightness and gamma settings are consistent. As monitors improve, gamma settings should remain constant, but do not assume that this is the case. A simple calibration program found in image processing software such as PhotoShop can help to do this calibration. Establish the settings for contrast and brightness at the beginning of the project within the adapted space and make sure all operators adhere to them. A set of guidelines with the appropriate settings will ensure the consistency of the digital images. Data about the settings of the capture device should appear as part of the metadata associated with the digital file. For more critical work you should use a spectrophotometer and calibration software such as the Eye-One Monitor (see http://www.gretagmacbeth.com).

Calibrate the equipment used for the digital process regularly. Most flatbed and film scanners can only be calibrated by their manufacturer because calibration is done at hardware rather than software levels. High-end digital cameras may have to be calibrated as the focal length and lighting may change with daily use. We recommend that calibration settings be defined at the start of each batch. Whether the work is being done in-house or outsourced, an appropriate policy or contract arrangements should be in place to ensure equipment is regularly recalibrated.

Color management

A Color Management System (CMS) should be employed by any institution wishing to accurately reproduce color from the source material (original object, artifact or film master) throughout the entire chain of digital hardware devices including monitors, dye-sub and ink-jet printers, film recorders and printing presses. A CMS is a group of software tools and hardware measurement instruments that work together to map the wide gamut of the original color space into the narrower gamut of display and paper output so that a reasonable and perceptible consistency and quality is maintained.

These systems are complex, evolving, and beyond the scope of this Guide. However, there is a growing need, both in terms of economics and image quality, for institutions to develop expertise and deploy a CMS. While lithographers, pre-press experts, image capture specialists and web designers all may use some form of a CMS they often are grounded in different nomenclatures. Therefore a standard form of communication is required to achieve the desired results in image distribution.

Fruitfully, there is a growing body of literature on the subject. The lingua franca is currently found in Adobe

Photoshop version 7, which leverages Apple's Colorsync and Windows ICM operating system core color technologies to provide cross platform color management. These systems use "ICC profiles" that are based on rules made by the International Color Consortium for storing color information.

The most accessible writing on the subject has been by Bruce Fraser who is preparing a mass market reference on color management for autumn 2002 release.[2]

Link Box:

Color Management Web References:

International Color Consortium (ICC): http://www.color.org/

Apple Corporation's ColorSync: http://www.apple.com/colorsync/

X-Rite: http://www.xrite.com

Munsell Color Science Laboratory: http://www.cis.rit.edu/mcsl/

Electronics for Imaging: http://www.efi.com/

GretagMacbeth: http://www.gretagmacbeth.com/

Targets

Targets provide a mechanism for benchmarking the capture process. In simple terms, a target is a sample with known characteristics that can be used to establish a baseline or standard by which to assess capture. Your project should adopt a policy on what targets and scales to use, when to use them, and how their quality will be controlled. Two types of targets are commonly used: edge and detail (resolution) targets and color or grayscale charts. See Resolution Targets Example Box in Section VIII: Quality Control and Assurance.

Ensure that the targets used are appropriate to the material being scanned, as they are different for transparencies, prints, and certain kinds of reflective material. You should use new targets at the start of your project as they lose their accuracy with age. Scan the appropriate target at least once a day, at the start of a new batch of material, or if the settings of the equipment are changed for any reason. Although some think that targets should be attached to every image, this is not possible for non-reflective formats and will have resource implications as images then have to be cropped for display (though it is sometimes possible to automate this process). Including a scale with each digital image can be of value. If it is not in each image it may be sufficient to include it in the first image of a batch and whenever the settings of the equipment are adjusted (e.g. the position of the camera).

Metadata

Metadata associated with digital objects is the key to their sustainability. Metadata assists in the management of the digital data, documents its origin and specifications, and assists with discovery and retrieval. For detailed information on metadata, see the Appendix on Metadata. Briefly, from among the many types of metadata there are three that are generally used with digital images:

- Descriptive metadata describes and identifies information resources, and is used at the local level to enable searching and retrieval and at the web level to enable users to discover resources. Examples of descriptive metadata include Library of Congress Subject Headings, Categories for the Description of Works of Art (CDWA) (http://www.getty.edu/research/institute/standards/cdwa/), the Art and Architecture Thesaurus (AAT) (http://www.getty.edu/research/tools/vocabulary/aat/), and the Dublin Core metadata set. The title and creator of the object are examples of descriptive metadata.
- Structural metadata captures information about the structure of a digital object and the relationships between its parts, for instance between page images and text transcriptions. It facilitates navigation and presentation. Examples of structural metadata are included in the METS standard (http://www.loc.gov/standards/mets/) and SMIL (Synchronized Multimedia Integration Language)

(http://www.w3.org/TR/REC-smil/).
- Administrative metadata facilitates short and long term management and processing, records data on creation and quality control, rights management and preservation. See Cornell University's "Web Hub for Developing Administrative Metadata for Electronic Resource Management" (http://www.library.cornell.edu/cts/elicensestudy/home.html). Another very useful resource is the NISO IMG draft standard on metadata requirements for digital still images: http://www.niso.org/standards/dsftu.html and http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

A number of issues arise when designing a metadata model for a project and these must be addressed at the outset if you are to ensure consistency:

- Static and evolving elements: will the metadata item need to be changed frequently, or is it constant for the whole project?
- Manual and automatic creation: many systems have automatic metadata capture (for technical elements) but some metadata may have to be created by the operator during the capture process. You will need to assess the importance of this manually created metadata and its role in your collection and delivery plans. Keeping manual metadata creation to a minimum will help reduce costs. But for some projects, detailed metadata created by experts will enhance the project's value so greatly that the cost will be more than justified.
- Who captures the metadata? The operator or an administrator? Very often this task is split between two or more people, particular if subject expertise is required.

Projects, on the whole, use a mixture of automatic capture and manual capture by the operator, subject specialist, or cataloger. The level and depth of metadata captured vary from project to project according to resource levels, including both staffing and equipment.

The operator of the digital camera or scanner will usually capture those technical metadata (as part of the administrative metadata set) that are not automatically captured by the scanning software. Descriptive metadata, particularly those relating to the original object, should be prepared in advance by a staff member with specialist knowledge of the collection.

File naming conventions

Before beginning a project you should define a file naming protocol so that your filenames will be consistent and intelligible. The filename can be part of the metadata and the information in the filename can be used to reflect the collection details, e.g. collection name, book title, or page number. One of the most common pieces of metadata in file names is an ID number. In file names you should always ensure that the three letters after the period are reserved for the file type information (.tif, .jpg) as this is important metadata for the operating system to know which applications can handle this file type.

Images of Text

Although the capturing and encoding of text have been dealt with in the previous section, here we will raise some of the issues in capturing pages of text as images, rather than as encoded text.

As already indicated, text can be captured in three ways: as an image only; as fully transcribed text; or as fully transcribed text with an associated image. There are currently no viable technologies for performing text searches on images, so if searchable text is required, some sort of transcription will be a necessity. However, for many purposes the original presentation of the text is also important, in which case an image of the text may be a useful addition. A source that is visual in nature as well as textual, for instance, an artist's book composed of both image and text, is best represented as both data types.

Example Box:

Projects that link images and text include:

- MOA (Making of America), using MOA2/METS: http://moa.umdl.umich.edu/

- Electronic Text Center at the University of Virginia Library: http://etext.lib.virginia.edu/
- American Memory at the Library of Congress: http://memory.loc.gov/
- Online Archive of California: http://www.oac.cdlib.org/
- AMICO: http://www.amico.org/

Projects that have only captured images of text because of the condition of the original material (for example, manuscripts or very early printed books), often store the best quality images so that when OCR technology improves sufficiently, or when more resources become available, they can create machine-readable text from these images.

Although proprietary file formats are generally not recommended for digitization projects, especially for archival versions, the PDF (Portable Document Format) can be very useful for a deliverable page image. You will have to buy the software to create PDF files (Adobe Acrobat), but the software to view PDF files is provided free (Adobe Acrobat Reader). The ability to create PDF files from any application and PDF's platform-independence make it particularly useful for delivering text material over the web.

Definition Box:

Portable Document Format (PDF)

- A proprietary format created by Adobe, but licensed to other software developers.
- Combines a precise page image (including fonts, graphics, and links) with a searchable full-text version of the document.
- Content can also be saved in RTF (Rich Text Format) for re-purposing.
- Users can read PDF files using the free Adobe Acrobat Reader, but creating PDF files requires the purchase of Adobe Acrobat.
- Provides only limited retrieval and representation of document structure.

Post-creation

Quality control and assurance is an integral element of the creation process, and the best digital surrogates should be created to minimize the work involved (see Section VIII on Quality Control below). Standards, technology, staffing and financial resources are all part of this process. Of all of these factors, perhaps the most significant is operator skill and training; investment in both of these areas will pay off significantly in the long run. However, there are also post processing techniques that may be needed to enhance the digital image. This is a practice carried out by a number of projects and generally in batch mode. The standard program used is Adobe PhotoShop© that has various filters and tools to assist in post image processing. Tasks commonly performed after the capture stage include unsharpen mask, gamma correction, noise reduction, deskewing, and cropping. Although you can do a lot in an image manipulation package, it is no substitute for care and attention to best practice at capture time. It is more efficient to set the guidelines for capture and make sure the operators adhere to these than to run image processing to redress errors made in the capture process.

The dynamic range (or density range) of a scanner is a secondary, but important consideration. If a scanner has a high dynamic range, it can sense a wide range of light levels. Dynamic range is usually higher in drum scanners and slide/film scanners. The dynamic range does not actually indicate how many different levels can be resolved but determines the smoothness of transitions between adjacent tones.

How scanner and camera manufactures implement this technology varies. A table in the Appendix on Equipment indicates the types and uses of scanning equipment available.

Conclusion

The successful digitization of still images begins with a careful examination of a set of complex issues. From selection of materials to the best means of capture and storage, these issues must be carefully considered for

each project. Questions the project manager asks may include the following. Which materials will be selected for scanning—originals or derivatives? If fragile originals are digitized, what safeguards should be in place to protect them? What file formats are most appropriate for the material? What are appropriate file sizes and what will be the impact on storage requirements? What is the best way to develop a file-naming protocol? What about sustainability? This section has answered many of these questions or provided a framework within which to begin to seek answers. For further insight, the reader should consult Section VIII on Quality Control and Assurance. That section further clarifies procedures that should yield a quality product.

---

[1] That said, it is worth bearing in mind that Manfred Thaller's work at Duderstadt (http://www.archive.geschichte.mpg.de/duderstadt/dud-e.htm) and work commissioned by the Library of Congress (http://www.loc.gov/) concluded that for manuscripts except under special circumstances grayscale scans provided sufficient information.

[2] Bruce Fraser, Fred Bunting, and Chris Murphy, Real World Color Management (Peachpit Press: forthcoming, autumn 2002). 600pp; ISBN: 0201773406.

[3] For digital cameras this excluded consumer digital cameras that tend to use Contact Image Sensors (CIS) for weight and cost considerations.

[4] This description relates to standard (Legal sized, 8½ x 14") flatbed scanners. Outsize (or large format) flatbed scanners are also available and employ similar technology. Sheet feed scanners employ an optional attachment to a standard scanner to enable automatic batch scanning of loose-leaf material.

[5] Engineering Scanners are also referred to as Wide Format Scanners.

table of contents previous chapter next chapter

valid xhtml 1.0 strict
abp~03/03

4 The Online Archive of California

HATII interviewed Robin Chandler, Director of the Online Archive of California, on January 26 2001. A fundamental facet of the California Digital Library, the OAC aims to make accessible to a wide audience, the vast array of primary sources and artifacts stored in institutions throughout California. Robin Chandler has championed the advancement of the OAC as an OAC working group member and as the Chair of the UC Archivists Council.

4.1 Organizational Digitization Program and Policy

" A core component of the California Digital Library (CDL), the Online Archive of California (OAC) is a digital information resource that facilitates and provides access to materials such as manuscripts, photographs, and works of art held in libraries, museums, archives, and other institutions across California. The OAC is a California-wide digital archives that integrates in a single searchable database of finding aids to primary sources and digital facsimiles of selected content from archival collections dispersed throughout the state. Critical issues range from ensuring that policies, procedures, and mechanisms are in place for contributing digital finding aids to the OAC database to coordinating the deployment of select digital content and the development of documentation, tools, and training needed to advance the growth and management of the OAC. The California Digital Library (CDL) is a "co-library" with a focus on digital materials and services. A collaborative effort of the ten University of California (UC) campuses, organizationally housed at the UC Office of the President, it is responsible for the design, creation, and implementation of systems that support the shared collections of the University of California. Several CDL projects, including the OAC, focus on collaboration with other California Universities and organizations to create and extend access to digital material to UC partners and to the public at large."[3]

As part of the OAC, there are a variety of CDL sponsored virtual archives sub-projects that in turn have project partners, such as the Museum Online Archive of California (MOAC). The MOAC project, comprising eight museums in California, is testing the application of Encoded Archival Description (EAD) to describe and provide access to museum collections, test specifications for item level description of museum artifacts and to integrate digital facsimiles of museum collections with digital archival materials as part of a union database. While hosting the collections of other institutions, the OAC creates virtual archives, such as MOAC and JARDA – the Japanese American Relocation Digital Archives. The material from JARDA comes from archives, libraries and museums from across the state. Currently, the OAC does not conduct a collection survey as such, as it does not have an inherent collection, but provides the platform for digital resources across the state. For example, individual OAC partners are creating virtual archives for the OAC, such as the Free Speech Movement, the Cased Photograph collection and the San Francisco Earthquake and Fire. However, one of the charges of the OAC Steering Committee – the database's administrative group – is to advise on the development of an overall collection policy for the OAC.

The priorities for selecting material vary from project to project. Research and public demand are major drivers, e.g. the JARDA project is in response to the OAC partner institutions' recognition of the high public demand for reference access on Japanese internment and this need drove the establishment of the project. Political awareness plays a part in the selection of material to ensure the collection is representative of the need for the specific resources and that it can be used by all possible users, with filters down to education, at levels K–12. The curators from the eight participating institutions formulated the collection development criteria used to select materials for the JARDA project. They determined that it was important to have balanced perspectives, that is, to provide the government's viewpoint (the camp administrators and the context of evacuation from homes) and the views of the internees, as well as to document life in all of the camps.

The OAC database has provided the first opportunity for geographically dispersed finding aids, encoded in EAD, to be available at a single online source in California. The next steps for OAC are to improve the database interface by working with user communities (faculty, students, the K-12 audience) and to improve access by identifying what functionality is required, what content is required, and how the interface can promote ease of

use.

The OAC is a collaborative venture; co-operation is an integral part of all projects, with materials from dispersed collections and diverse institutions. This collaboration is worthwhile and effective but can often take more time than may at first be thought. Partners contributing finding aids and digital images to the OAC must adhere to CDL standards, such as the Digital Imaging Standards and the OAC Best Practices for New Finding Aids. Effective tools and services can be built using standardized data.

While the nature of the source material varies considerably from project to project, most forms are digitized. The OAC is currently reviewing available standards for incorporating the presentation of streaming audio and video datafiles into the database.

Standards are an important issue to OAC to ensure the access and longevity of the diverse materials. To describe content, they use MARC, EAD, TEI Header and CDWA. During the next year, it is anticipated that a planning group will be established to develop guidelines for incorporating non-EAD descriptive metadata standards into the OAC. Standards have been used where possible, and where standards are not used, it is because they do not exist as yet. Standards are also used for for data control values – AAT, and Library of Congress Subject Headings. They do not reject standards, nor adapt them. The union catalog allows information sharing. There are different data sets but generic tags. This bears the most basic level of tagging and allows diversity from the equally diverse collections.

4.2 Project Management and Planning

There are formal project management structures in place within the OAC and projects are attached to these. There is a Steering Committee that is the principal advisory group of OAC, responsible for evaluating issues of administration, intellectual property concerns, funding resources, and content development. The OAC manager has direct management of the various projects, such as JARDA, MOAC and California Cultures. The OAC manager is in contact with forty institutions and communicates with the OAC Working Groups advisory panel which directly oversees several subcommittees: Architecture & Structured Text, Metadata Standards and Digital Objects. This management structure allows flexibility and communication from project to project and from the strategic top down. The OAC manager fulfils a pivotal role as the point of contact for all parties.

The individual projects have various project management structures, since they are each led by an individual Project Manager. Each will have an advisory board that helps and assists where necessary. The OAC manager liaises with editorial boards and the project managers on each site and campus, so the levels of communication are established and transparent to all. Targets are set and a timeline established. There are feed back loops that enable the communication to flow freely and where necessary staff will be redeployed to reach a goal or target. The OAC Manager is considering using project management software for all tasks; they currently use Gantt charts and the present project management structure has experience in this practice.

The OAC is a core component of the CDL and is firmly established in the management structure. As a part of the larger body of the CDL, the OAC has its objectives clearly established. The ability to interact with the other components of the CDL is an essential part to establishing success. The CDL web page describes the organization structures:

“The CDL effort is directed by the University Librarian and Executive Director. There are separate staff responsibilities for Shared Content, Business Development, Education & Strategic Innovation, Scholarly Communication Initiatives, Digital Library Services, and Digital Library Technologies. There are several cross-campus advisory bodies to the CDL. The CDL maintains liaisons to all the UC campuses as well as to its other partners such as the State Library, the California State University system, and other organizations throughout California”[4]

The OAC is placed within the Shared Content section of the CDL. There are broad areas within this structure that allow for collaboration and sharing of skills and expertise.

While pilot projects and feasibility studies are not generally carried out within OAC, projects may do this independently, however, it is neither a guideline nor is it advised.

Currently, the OAC carries out its digitization in house using a variety of technologies depending on the projects. However they are now considering outsourcing for this process due to concerns about quality control that have emerged from the in-house digitization process. The OAC is in the process of interviewing vendors with attention

to quality of product, considerations for handling rare and fragile materials, and security for materials. Longevity and access are the main drivers of the CDL and OAC and therefore the digital objects must be of the best possible quality. The funding is mainly public money and thus there is a large level of accountability and need for the digital materials to be available to the public in excellent quality for as long as possible. The use of outside vendors implies that this accountability can be shifted from in house staff to these vendors. Not all in-house staff have the same levels of skills and there are limitations to the equipment used, thereby raising some issues of quality control. The contract with vendors enables the quality control levels to be established at a set cost that will not shift if there is a skill differentiation.

This strategy is to contain funding by establishing a contract that is creative and collaborative. All aspects of the process must be covered by the contract with quality control checks and benchmarks established clearly with minimum targets. Post processing must be approached in the same manner.

Outsourcing is a viable option and shifts the quality control and workflow to the vendor while ensuring that the cost will not rise as a result of re-imaging where poor quality imaging has been discovered.

4.3 Project Life-Cycle Processes and Procedures

The copyright is owned by the institutions contributing the digital images, which is clearly established where the original archival materials are publicly available. OAC is developing a model for indicating rights with the digital objects, similar to the Bancroft Library statement [see 2501AG09]. Different project and archives have statements that apply to each but the strategy is to create a model for stating IPR on all archives/projects covering as much ground as is possible. The diverse nature of the materials, in terms of both content and geography, will mean that the copyright will vary.

The California Digital Library Technology Architecture and Standards Workgroup (TASW) is responsible for recommending to the University Librarians and other appropriate University bodies for review, architectural guidelines and standards for the University of California shared digital collections. These guidelines should be consistent with industry and other University standards, provide a framework that facilitates the creation of integrated systems and provide the flexibility to foster innovations in scholarly communication. As such TASW has developed standards for architecture, digital imaging, best practices for image capture and digital image repository models. The digital image repository model calls for storage of masters and derivative images at both CDL and partner institutions. The partners maintain control of their master image and the CDL master will serve as a records management copy. The OAC Operations Group, the technical group comprised of computer programmers and the OAC Manager, is charged with developing and implementing the storage system for managing digital images housed at CDL. CDL oversees the accessibility, management and storage of EAD and MOA2 metadata. Derivative images are primarily stored by OAC, and users navigating EAD metadata primarily access images stored at local and not CDL servers.

The OAC has EAD finding aids for forty separate institutions, the majority of which are not part of the University of California (UC). The OAC Working Group Subcommittee on Metadata Standards is currently conducting a survey to determine how many of the 5,800 finding aids in the OAC also have a MARC record. This survey is the first step towards developing a plan to create MARC records for each of the finding aids in the OAC. Most of the UC campus library finding aids do have MARC records that are available through the Melvyl union catalog (the online library catalog for the UC system). The MARC records contain an URL that links to the finding aid in the OAC database

The OAC co-ordinates EAD creation for the finding aids for the digital objects. The OAC Operations Group is continually working towards the further development of procedures for delivering new and maintaining existing finding aids and is currently discussing the formulation of a quality control tool for OAC EAD based on the “best practices” and metadata standards for processing new finding aids in EAD. The OAC also identifies or develops further tools to facilitate EAD mark-up of new finding aids for the digital objects.

As each project creates the metadata for each digital object, e.g. EAD, TEI Header and MOA2 (see Bancroft report, 2501AG09), the OAC extracts information for the finding aids for online access. The in-house method chosen by the OAC to catalog its own digital archives is an Access database and File Maker Pro. All the forms of metadata are described, the original object, the digital object, the digitization process, technical details, staffing details and administrative information. The digitizer will record the details about administration and structure while an archivist will record the descriptive metadata. The metadata catalog and the main catalog are linked by the MARC 856 field.

4.4 Quality Control

OAC has reached a point in its development where it is reviewing the digital imaging process and recognizing the need for formalization of the quality control process. While the CDL TASW has developed Best Practices for Digital Image Capture, there are no formal quality control procedures in place. This is also a leader to the outsourcing ventures, as the vendors have logarithms for skew and understand issues such as color space, dynamic range and gamma correction, which until now were dealt with uniformly by the institutions. This may have had the effect of creating images that are not as high quality as is possible. This reflects the need for photographers to be involved in the digitization process, as they have the skills and knowledge required to understand and create the best images, whereas cultural heritage professionals with some knowledge of IT are not always the most experienced or knowledgeable to carry out the physical digitization. The level of quality control required will depend on the project, the volume of material and the practitioners involved.

4.5 Access to Digitized Materials and Evaluation

Currently, access to the digital objects is provided by drilling down through the hierarchy of EAD finding aids available in the OAC. The OAC is now exploring the creation of a database comprised solely of digital objects so that the user could either access images through browsing and searching finding aid data, or directly by accessing an image database. This is a dynamic process; scholars rely on navigating the finding aids to learn about collections and review digital objects. But the complex finding aids may not be suited to the multiple users across the community who are unfamiliar with primary resources. Collaborating with CDL's eScholarship program, the OAC is exploring means to enhance access to finding aids and digital objects available through JARDA. A group of faculty, K-12 educators, digital librarians and archivists have been identified and are meeting to discuss interface development, interpretive tools required, collection development strategies and metadata standards that will improve access to the OAC and its virtual archives resources.

At the same time, as XML becomes more integrated and internet browsers are supporting it, the CDL is developing a plan to convert the SGML EAD finding aids to XML and concurrently identifying new software to succeed Dynaweb. The OAC would like to improve upon the efficiency of the Dynaweb software, and take advantage of XML tool development targeted for browser deployment.

---

[3] http://www.cdlib.org/about/faq/units.html#collections

[4] http://www.cdlib.org/about/faq/#organization

valid xhtml 1.1
abp~04/02

14 Digital Imaging and Media Technology Initiative (DIMTI)

On January 8 2001, HATII interviewed Beth Sandore, Professor of Library Administration and Head of the Digital Imaging and Media Technology Institute at the University of Illinois. The primary objective of DIMTI is to create digital deliverables as resources for teaching, learning and research. The Institute has liaised with other organizations both nationally and internationally, with a view to creating a fully comprehensive archive. While the intended audience for these resources is predominantly the University community, the Institute recognizes that audiences outside this sector can also benefit from access to the material, and continues to make developments in this area.

14.1 Organizational Digitization Program and Policy

The University of Illinois's Digital Imaging and Media Technology Initiative's (DIMTI) digitization activities began in late 1994 when the University of Illinois Library was approached by two corporate sponsors about academic digital imaging. Kodak wished to promote its PhotoCD format and the Follet Corporation, better known as a textbook distributor, was interested in collaborating to develop image databases through the digitization of special collections. Immediately following this the Library joined the MESL project at its outset. This provided a ready made database that familiarized the Library with issues related to large-scale repositories, some 10,000 images over 4GB of space. The two years of this project were the most useful and productive for DIMTI and entailed agreeing upon issues such as intellectual property, metadata and interface design amongst the seven universities involved. This project also carried out user surveys with 350 students and 20 faculty members, and Getty funded several very useful MESL meetings. Although there was no formal collection survey as part of this project the Library interviewed local curators, special collections staff, archivists and librarians and made a call across faculty to suggest projects. The results of this call were also used for the Kodak PhotoCD project mentioned above. The Library learnt about guidelines and standards and suffered for a lack of these. As a result, they developed their own metadata standards early on, and have subsequently adopted a number of metadata image file format guidelines.

Involvement in these projects has helped establish priorities for digitizing holdings. A formal strategic policy statement was developed in fall 2000, with initial objectives to be realized by 2003. The responsibility for the creation of this statement lies with the Library Information Technology (LIT) Committee, which advises the University Librarian and Library faculty about emerging information technologies. This policy has been approved by the University Librarian and two advisory groups (an Administrative Council and Executive Committee). The Library has a very flat administrative structure and these two bodies help build consensus across the Library. The establishment of digitization objectives by the LIT is the most tangible problem in the development of this policy and while most are optimistic about moving forward in this fashion, some feel that the Library needs a hierarchical structure, in particular with regard to creating a vision.

The Library is also involved in the University's Advanced Humanities Technology Group, which discusses the application of advanced technology and has aided digitization initiatives on emblem books, historical maps and aerial photography. At present, the DIMTI has adopted a set of rolling goals that are revised yearly. In practice the DIMTI does not undertake a project unless there are faculty who want to do it or will use it.

One major obstacle to the development of a strategic policy has been that the institution has not fully addressed the need for an internal re-allocation or new funding, as it cannot rely solely upon external funding. A recommendation to other institutions that are attempting to build digital library programs would be to plan the infrastructure so that digitization operations become integrated into mainstream activities, although this may be implemented in different ways.

The main overall obstacle to planning the development of digital deliverables from the Library's collection has been how to deliver these. A strong thread through DIMTI has been the undertaking of exploratory projects, but this has been relatively easy. The difficulty has been in determining intellectual access in terms of appropriate benchmarks and resolutions. Therefore, an obstacle to the process of building digital deliverables has been the

need to develop robust databases and structures, and testing metadata and interoperability. DIMTI would strongly suggest that if you don't know how you are going to deliver material, do not digitize it. To help in this process, DIMTI have under development a software applet that determines capture settings according to type of material, size, etc., using the Cornell quality index formula.

As a result of the above deliberations, DIMTI's primary selection criteria are the material's teaching and learning potential, as expressed through faculty interest, which in turn contributes to new research content. Another primary selection criterion is the material's historical and cultural value expressed in terms of uniqueness, an example of this is the Library's emblem book collection. Secondary selection criteria are enhanced access and preservation (against content loss). DIMTI places access criteria in the context of usage and scholarship. For the digitization of aerial photographs, preservation was the number one priority and here the cost projection of $40-$50 per image was not as expensive as cataloging a book. These cost projections represent one element of a set of further considerations for digitizing any collection: Does DIMTI have the tools? What is the cost? Do we have the expertise? What is the cost of metadata? All these criteria have stayed stable since 1994, however, DIMTI believes that digitization for archiving is going to come into the foreground.

The DIMTI has co-operated with other university archives, libraries, museums, historical societies, academic faculty and users and corporations at the local, regional, national and international levels. There are also a number of projects planned with further co-operation in mind.

The DIMTI has no anticipated end date, but the Head suggests that they may not exist in ten years' time if the goal of integrating digitization activities into the mainstream activities of the Library is achieved. In this scenario there may be just one coordinator to provide an overview of the various projects and ensure they are managed in a complementary fashion.

The primary purpose in creating the digital deliverables reflects DIMTI's selection criteria. Digital deliverables are primarily a teaching, learning and research resource. Revenue generation was a factor injected at the start of the initiative but has subsequently been put aside. However, DIMTI is beginning a small experimental project to publish some selected parts of collections with the aim of cost recovery. This material would also be made freely available by the Library on the web.

The DIMTI produced a two-paragraph mission statement, published on its website regarding the significance of its activities.

The type of source material digitized includes:

- printed books, serials and similar bound material (published);
- printed documents (gray literature);
- unbound printed documents;
- typescript documents;
- printed illustrations;
- handwritten documents;
- artifacts;
- original artworks;
- intermediaries;
- hand copies;
- photographic negatives;
- photographic prints;
- sound recordings;
- video recordings.

The predominant medium of the materials to be digitized is paper, but for visual, graphical material the form varies from 6x7" to 3x4' for oversize maps. Until now the digitized deliverables have usually represented a sample of material rather than an entire body. For the DIMTI the intention to re-purpose the digital deliverables has always been part of its remit. For example, they work with the educational technology group on campus and elementary school teachers to produce electronic guides using some material and metadata. This approach has worked well and the DIMTI has found that it has been most common for faculty members to point to material for web syllabi.

The following standards, guidelines or tools are used for representing content:

- EAD
- TEI;

- CIMI;
- Unicode;
- JPEG, GIF TIF;
- MPEG;
- SGML;
- XML.

The following standards, guidelines or tools are used for describing content:

- MARC;
- Dublin Core;
- EAD;
- TEI;
- Categories for the description of Works of Art (CDWA);
- VRA (Visual Resources Association) core categories.

The following standards, guidelines or tools are used for controlling data values:

- AAT Art and Architecture Thesaurus (Getty);
- Library of Congress subject headings (LCSH);
- Thesaurus for Geographic Names (TGN—Getty);
- Union List of Artists' Names (ULAN—Getty);
- ICONCLASS;
- Library of Congress Thesaurus for Graphic Materials (TGM);
- Library of Congress Name authority files.

The program looked at other existing guidelines for digitizing particular document types when planning its digitization strategy. These included TEI recommended best practices at the University of Michigan, the IFLA Metadata Website, the Berkeley Digital Library SunSite, and the Library of Congress National Digital Library Program metadata sites. The DIMTI modified Dublin Core for a few projects where documents had to map to the basic metadata scheme held in the project database.

The following standards, guidelines or tools are used for representing structure:

- HTML;
- XML;
- RDF.

In relation to standards in general and navigating between the ideal and the realistic, the DIMTI has drawn from the ideal metadata in the Making of America 2 Interim Report and found that providing at least 15 Dublin Core elements has worked well to identify static images of visual materials and text documents.

The intended audience for the digital deliverables depends on the project but the university and research community is at the heart. However, there is also significant K-12 and general public use. With regard to metadata, DIMTI think in terms of the university community and then try and gear the metadata to K-12 where necessary. This has not been as hard as one might think.

As part of the MESL project, an evaluation of the university target audience was undertaken. This established a baseline across the seven universities involved, for preferences in functionality and access to an art image database. This demonstrated that rich metadata was as important to users as the images themselves. In most cases the profile of users has been that anticipated, except for heavier general public use.

To an extent, DIMTI has recognized the needs of those with disabilities, but there are no text-only versions due to the exploratory nature of the projects. In only one case are there limitations to the use of the digital deliverables where access is restricted by IP address. This is a subscription to the AMICO database.

14.2 Project Management and Planning

Both internal and external advice was available on managing the digitization program. Internal advice included using the Survey Research Laboratory for statistical analysis and the Division of Academic Outreach Web Technology Group on campus for advice on the design of instructional material. External technical consultancy was used for aspects such as interface design and search engines, although this tended to be informal. The

existence of DIMTI has led to changes in organizational relationships and procedures. The involvement of the LIT Committee in the development of a strategic plan is a recognition that DIMTI and other projects are viable and provide a locus of expertise. An ineffective project management procedure has been attempting institutional collaboration without ensuring the partner institution has sufficient human resources in place to support the work. Managerial quality assurance procedures include developing timelines and plans in advance, with specific milestones and events. These plans are revisited every quarter and made available on the web for all parties.

Feasibility and pilot studies have been carried out for scheduling, training needs, technical feasibility, user needs, workflow analysis, workflow piloting and technology forecasting. DIMTI have made changes to the design of projects as a result of preliminary studies. The projects have not departed from their original goals but the timelines have been changed based on technology changes and the ability of partners to deliver aspects of the project.

Benchmarking studies are undertaken, which include determining network throughput, however, all procedures are analyzed throughout. The delegation of work is determined by externally funded projects, where a project coordinator and dedicated staff are hired. There is more flexibility in internal work, where staff have been hired for their areas of expertise and to develop the skills set of the core staff.

Some digitization has been outsourced but there has not yet been a large enough project to justify this. In-house digitization has been the predominant method, since the student body represents a captive and cheap labor source and the University of Illinois Printing Service has a Xerox Digipath that can handle batch digitization of text documents. All other equipment has been bought in, as there was nothing available at the start. The decision of which digitization process to adopt, depends on the type and volume of material and the conflict levels involved. For example, if the material is rare or original it is currently more acceptable to set up the digitization on site. For its digitization capture, DIMTI uses two Epson flatbed scanners for material up to 12x17” (800dpi, 36 or 48 bit color input/output), a large format color scanner, a 1500dpi color slide and film scanner and a 3 megapixel Nikon D3 digital camera. DIMTI also has a large format six-color inkjet printer and a 600dpi color laser printer. Guidelines for data capture procedures for this equipment have been established. Benchmarks used for image digitization are gray scales, color charts and reproduction charts. The slide scanner is calibrated using a printer color palette and the digital camera will use the vendor's calibration software as soon as it becomes available for the Windows platform.

14.3 Human Resources and Training

The unit employs three FTE professionals, including a full-time director, and two Visiting Special Projects Librarians who co-ordinate externally funded projects. It also employs one FTE graduate student who specializes in programming and database development, digitization benchmarking, and web page creation. The professionals have expertise in metadata, digitization, database design and development, instructional materials development and evaluation. For professional positions, staff have a library degree or equivalent degree experience. Graduate student employees have some interest in library studies or information retrieval. In addition, various Library staff and faculty participate in specific projects in a team environment, and receive necessary training.

Training needs are assessed through talking to project team members and asking about their skills. Areas of training that have been identified include:

- project management;
- application of technical standards;
- preparation and handling of materials for digitization;
- technical operation of digitizing equipment;
- post digitization processes e.g. editing;
- metadata creation;
- digital preservation.

The team members that have received training include:

- project director;
- curatorial staff;
- specialist technical staff;

- library/cataloging staff;
- equipment operators.

Training has been organized using the following:

- in-house, using project staff;
- in-house, using external consultants;
- attendance at external courses;
- independent study;
- learning on the job.

The training has met the needs of the project.

14.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and owns the copyright in most of the original materials. The copyright status of the digital deliverable is declared in a statement on the project's website. DIMTI is comfortable with the effectiveness of this practice. Material in copyright was digitized with the owners' agreement. Users are allowed to make printouts of the digital deliverables on paper and download to a PC or LAN.

For textual material users can download:

- ASCII text files;
- TEI or EAD marked-up text;
- XML marked up text;
- PDF;
- HTML marked up text.

For digital image material users can download:

- thumbnails;
- reference and screen-size lower quality images;
- associated documents.

For digital image material users can only view highest quality images on-site.

For digital audio material users can download:

- less than 30 second samples;
- full length compressed;
- associated documents.

For digital audio material users can only listen to high fidelity sound on-site.

For digital moving images users can download:

- samples;
- lower quality digital video clips;
- associated documents.

For digital moving images users can only view highest quality digital video clips on-site.

The Digimarc electronic watermarking management system is in use for selected projects and this has proved effective.

DIMTI recognizes the need to develop a conservation procedure for original materials. The project works with curatorial staff to assess the condition of original materials prior to digitization. At present there is no head of preservation in the Library so conservation activities are incomplete. Previous work has included encapsulating material and ensuring that materials are stored in acid-free sleeves and boxes. An example of some of the problems identified prior to digitization was the aerial photography collection. These had not been stored in ideal conditions and the photographs had become curled, the emulsion cracked and there were drawings and marks on the prints. Procedures to minimize risks include using the gloves for handling, acid free photomounts and

limiting light exposure. DIMTI has used cool lights but the deliverables were not as good; they have also stayed away from digitizing fragile books because they do not have a cradle. Materials are prepared and monitored by curatorial staff during digitization. Access has not yet been restricted to any original material because of digitization. Ultimately this depends on the fragility of the collection and the Library is conscious of its role as a public research institution.

Much of the original material in special collections is uncataloged or has a finding aid but most material is cataloged at the collection level in the Library system using MARC records. DIMTI uses all the information that is available from these records during the digitization process. The project sometimes must locate core reference material to complete the digital deliverables or alter material from its original form for digitization. Material has been rejected for digital imaging because of skewing, print skew or blurred pages.

Intermediaries (i.e. a second generation medium derived from the original) used for image digitization are:

- slides, 35mm or 4x5" transparencies;
- photographic prints;
- microfilm;
- single frame microfiche.

Intermediaries used for audio digitization are audio cassettes.

Intermediaries used for moving image digitization are video copies from film.

In only one case did DIMTI catalog original material; in this case an inventory list was produced in collaboration with the curator concerned. This work was written into the grant for the project. An institutional decision on how to catalog digital deliverables has not been reached yet, for example on whether or not to include details about the original in the metadata. In its cataloging, DIMTI includes information on the original on the basis that users first want to correctly identify and find material and then decide in what format they want to receive it.

The following standards or guidelines are used for cataloging the digital deliverables:

- Dublin Core;
- MARC;
- FGDC—Federal Geographic Data Committee (U. S. Geological Survey—for geospatial data).

DIMTI developed its own in-house system for holding metadata using off-the-shelf database and scripting software (SQL, MS-ASP).

Tools used for controlling data values are:

- Thesauri and other controlled vocabularies;
- Indexing and cataloguing guidelines.

Metadata details are recorded about:

- the original object;
- the digital object;
- the digitization process;
- technical details;
- administrative information.

Metadata records are created by either an information professional or the digitizer who is also an information professional. The decision of who undertakes this work depends upon the project, size of the workload, specificity of the metadata and the qualifications of staff. The non-MARC metadata records are held in a separate catalog from the main Library catalog and are available on the internet. The records in the two catalogs are partly independent of each other, with some similar features.

MARC and Dublin Core records for digital resources were created as part of the OCLC CORC project. An online catalog system was implemented three years ago, which can use the MARC 856 field, although the Library does not locally have a convenient capacity to use this feature at the moment.

14.5 Format, Resolution and Compression of Digitized Materials

The formats for retroconverted text-based digital deliverables are:

- SGML;
- XML.

No texts contained non-Latin scripts. OCR is not used a great deal as a conversion method for textual materials. It was used for scanning facsimiles and 65% accuracy was achieved (higher order ASCII character set). Initially Paperport suite was used for OCR and then TextBridge and OmniPage. The aim of using OCR was for automatic indexing, enhanced searching and computer based analysis. No special treatment was carried out prior to OCR. From this experience, DIMTI would recommend thinking carefully about the need for full text capability. DIMTI is looking at using Adobe Distiller for on the fly OCR to create indexes. No major "keying in" conversion has been done, only where it has been important to save the intellectual content at 100% accuracy.

For image material, the TIFF file format is used for the capture and preservation of master images and JPEG or GIF for delivery. Photo-CD has also been used for capture and JPEG or GIF for delivery. The capture and preservation resolution varies, depending on the original material and the benchmark. Delivery resolution averages 72-150 dpi, and is rarely more than 200dpi. Capture and preservation bit depth for bi-tonal images is 1 bit, for grayscale it is 12 bit and for color 24 or 36 bit. Delivery bit depth is 1 bit for bi-tonal, 8 bit or lower for grayscale and 8-24 bit for color depending on the material (for example continuous tone material is delivered at 24-bit compressed and illustrations are at 8 bit). Average file size for capture and preservation depends on the project but is in the range of 15-20 MB although with wide variation. Average file size for delivery is 200KB or less. JPEG image compression is used for delivery, to enhance usability. The project retains the original scans in uncompressed form. The project carries out post processing on images using PhotoShop for sharpening, unsharp and masking. For this process, testing on samples is undertaken on similar groups of material and scripts are then written for batch processing. Occasionally the contrast or color balance is increased.

For others starting work in digital imaging, DIMTI would recommend resisting the temptation to draw conclusions from limited experience – these will change as the scope increases. It is sensible to take on larger scale projects because common factors can then be seen. DIMTI also recommends using the Cornell guidelines as minimum standards (taking care that they are based on text scanning), while balancing these with local benchmarking practice and knowledge of users' needs and potential uses of the digital objects through pilot projects.

DIMTI has recently started working with sound digitization and uses WAV and RealAudio formats for capture, preservation and delivery. The audio sampling bit rate is 20 bit for capture and preservation and 16 bit for delivery.

The quality control procedures in place for the digital deliverables are a random set of checks on materials and a percentage check on carriers. Metadata quality control procedures are question and answer checks on the database, field by field, using human and script review. In addition there is scripted review of the web pages. Quality control procedures have become part of the daily routine work of staff. As a result, problems are identified and corrected along with the routine workflow wherever possible.

Users have open access to the catalog plus materials. Searching and browsing facilities are full text, fielded searches, simple and advanced modes with SQL employed server side. This same system handles metadata searching. Users are also able to save sets of selected items in some cases. The level of usage is in the region of 25,000 uses per month, monitored by automatic data capture and focus group interviews. Users do not have to pay to use the digital deliverables. Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, broadcast media coverage, conferences, meetings, email and conventional mail shots and registering with search engines. Press releases, registering with search engines, emails and conference presentations have been the most effective methods of dissemination.

14.6 Evaluation, Funding and Long-term Sustainability

For the MESL project, the DIMTI carried out front-end evaluation of users with a pre and post-development survey that included paper and online questionnaires, email, focus group discussion, user observation and computer interaction of logging. This led to changes in the interface and search options, metadata display and the underlying index.

Formative evaluation was carried out locally using focus group discussions and user observation, and this contributed to the changes mentioned above.

Summative evaluation was undertaken using paper and online questionnaires, email, focus group discussion, user observation and computer interaction of logging. This was undertaken to determine how the availability of digital images affected the established use of visual resources for coursework and research. If DIMTI were to repeat its evaluation work, it would follow up with focus groups. The main suggestion from the quantitative evaluation was that the more users learned about digital image availability, the less adequate they felt their skills were to effectively integrate images into their coursework and research. The results of this evaluation were disseminated through conference proceedings and are available on the web at the Getty Research Institute site.

DIMTI's budget is approximately $150,000 per year, excluding the Head's salary, and 75% of this is external funding. If DIMTI had had $200,000 a year they would have been able to hire a dedicated research programmer and a colleague to work on faculty liaison and content standards. The DIMTI's view is that in this environment the use of standards has saved money in the short and long term. Funding organizations' monitoring of DIMTI is built into the grants. Funding organizations have requested project management plans, cost models, workflow reports and user evaluation reports from DIMTI.

New material and metadata are added on an ongoing basis with periodic changes to metadata, user interface and file formats. DIMTI does not have a formal strategy for preservation to ensure long-term access. At present DIMTI checks data integrity annually and uses tape back-up and three sets of CDs stored in climate and light controlled conditions, but not cold storage. This is an informal stopgap strategy, based on original hardware and application software retention and migration of data, until the Library develops a formal plan.

It is intended to keep the digital deliverables available for at least 25 years, or such time as their use is not deemed sufficient to make widespread access available. Digital preservation strategies are currently being explored. The resource will not generate sufficient income to sustain DIMTI and the project does not have an exit strategy. This is one major problem that all staff are working on; a revenue stream needs to be identified from a combination of internal re-allocation, new money and external support. Loss of the digital deliverables would matter.

14.7 Conclusion

DIMTI is an excellent example of "joined-up thinking" in relation to digitization. The benefits of establishing a body such as DIMTI at the outset of digitization is demonstrated in the strategic overview that DIMTI enjoys. Whereas other libraries are dealing with formalizing policy they are doing so on top of largely autonomous, independent and entrepreneurial projects. Inevitably, trying to create a unified institutional vision or procedure is more difficult than if a single body had responsibility for dealing with generic issues from the outset. Furthermore, DIMTI demonstrates that such a body does not need to be large when supported by other appropriate administrative or managerial bodies. The fact that the University of Illinois Library has a relatively flat structure seems to have been beneficial in this case. Lastly, the goal of integrating digitization into the existing Library activities and structures commends itself for being both far sighted and an effective mechanism to garner institutional support and maintain digitization activities in the long-term.

valid xhtml 1.1
abp~04/02

8 Colorado Digitization Project

On January 22 2001, HATII interviewed Liz Bishoff, Director of the Colorado Digitization Project and Sue Kriegsman, the Operations Co-ordinator. The CDP represents the collaboration between several archives, libraries, museums and historical societies throughout Colorado State. It seeks to increase public access to the resources contained by these institutions, while ensuring efficient delivery of primary text and visual resources in digital form. The Project also intends to establish comprehensive criteria and standards for the selection of material to be digitized, thereby encouraging ongoing activity in this area.

8.1 Organizational Digitization Program and Policy

The Colorado Digitization Project is the organizational center for a number of individual projects throughout the state of Colorado. At the time of the interview there were 30 projects involving 48 organizations. There was no collection survey as such of the individual projects. However, the scope of the funding was primarily targeted at areas associated with Colorado – its scientific, cultural, and historical heritage – and those projects funded had collections that were pertinent to these requirements. For example, a project whose aim was to digitize coin collections from China, although valid and interesting in itself, would not be funded as it would not meet the CDP's criteria.

A survey was sent out to 300 possible projects to identify the content of the collections that might be digitized, and to those who had attended the December, 2000 CDP Leadership Seminar. Four subject areas were surveyed: Alpine Life (including flora and fauna, snow and floods); Water Resources; Colorado History; and Immigration/Migration. Based on this survey the CDP focused on more specific areas: water resources and trails. The subjects of the collections in the CDP fall into two distinct tracks: the historical societies/museum collections and the Library/archive collections. However, the CDP does not distinguish between different institution types in the guidelines, training, standards and funding. The content of the digitized collecitons has to be primarlily based in Colorado.

The CDP solicited projects from institutions including Denver Public Library, University of Colorado, and University of Denver. These are well organized collections and some had already started digitization, but the CDP's primary objective was to create new images – 50,000 in two years. Some institutions will digitise 1,000 items, while the larger grantees will digitise 3,000 items each.

Projects that wanted funding for software development for existing images were not accepted as this is not the primary remit of the CDP.

The CDP has a variety of working groups that help define standards and guidelines for the individual projects and a Collection Management Working Group operated as the grant reading panel. The content also has to be of interest larger than the local community – the State of Colorado and the areas identified as the primary focus are the drivers for the project and thus the resulting digital deliverables.

The collaboration between the projects, from the largest to the smallest and from the rural to the urban, is without doubt the most interesting and fruitful aspect of the CDP.

The main criterion for the selection of material for the overall project is scientific, historical and cultural value. Teaching and learning and intellectual property rights are important, as are research significance, provision of user services and improved functionality. Disadvantaged groups include the geographically diverse. Other issues are addressed at local level, such as conservation and potential commercial exploitation. The K-12 community is high priority, and teachers are motivated by students' need to have access to these resources.

Collaboration is one of the main objectives of the CDP, and this has naturally taken place with a variety of projects, including archives, libraries, museums, and corporations (in the form of public broadcasting stations examining archival video and moving the channels to digital). These collaborations are local and regional. At the time of this interview, the CDP was exploring the concept of a Rocky Mountain states collaboration project and

including nearby states, such as Wyoming, Kansas, Nebraska, New Mexico, Arizona and Utah.

The CDP has great experience in collaborative projects through the Director and through the willingness of the projects to be involved. Its success has been attributed to a commitment to equal participation, where all members are equal players. Everyone contributes to the concepts which shape the project. The steering groups and board reflect the four major types of institutions. In new projects started specifically for the CDP, there is a high level of camaraderie due to the flexibility which the minimum standards set by the CDP offer each project. Many of the smaller projects look to larger ones such as Denver Public Library for assistance, but the CDP recognizes that, while not all projects can perform at this level, they can nevertheless produce worthwhile material and run a successful project if they follow the standards and guidelines it sets. These standards were developed with this sort of collaboration in mind, so that the smaller projects can play an equal role along side the larger institutions.

It is necessary to have people with cross cultural heritage experience on any steering group so that one type does not dominate at the expense of others. The knowledge base must be wide, with representatives of different areas, for example metadata and collection management.

Communication is at the heart of successful collaboration. Email and conference calls are useful day to day tools, however, face to face and site visits cement the collaborative relationship at all levels, not only at management but also at operator level, which ensures inclusiveness for all projects, no matter how small or where geographically placed. The CDP runs regular and extremely popular training sessions on all aspects of digital imaging, and these have the additional effect of creating a community that knows and relates to each other as well as improving skills.

8.2 Project Management and Planning

The CDP's website describes the main goals and purposes for the creation of digital materials. This statement of intent is explicit about the scope, rationale and primary audience for the project:

"The purpose of the statewide-digitized Library is to work collaboratively to bring together, from all corners of the state, digitized materials that will:

- ensure public access to the rich resources in Colorado;
- promote the economic and efficient delivery of full text and graphic resources to the people of Colorado;
- contribute to the national effort to develop digital libraries and museums;
- work with educators to help meet the Colorado history standards."
- "To create an open, distributed, publicly accessible digital Library that documents crucial information for the residents of Colorado;
- to establish a collaborative structure among the state's libraries, museums, high schools and other organizations to coordinate and guide the implementation of the digital Library-museum;
- to establish criteria and standards to guide the selection of materials for inclusion in the digital Library;
- to demonstrate the value of libraries/museums in the emerging electronic information environment and their important contribution to the state's economic development;
- to assist libraries, archives, historical societies, and museums in the digitizing of materials and managing digital objects;
- to emphasize the content and rich resources held by Colorado libraries and museums." [5]

The source materials vary widely and include letters, diaries, government documents, manuscripts, photographs, maps, digital versions of exhibits, artifacts, and oral histories. The different projects determine whether the material is a sample or the entire body of resources.

The CDP has used standards and created guidelines for the projects involved to adhere to. These are minimum standards to enable the materials to be integrated into the union catalog, the key to accessing the resources from all the projects. The Project looked at CIMI, Dublin Core and other metadata standards and is exploring EAD. The guiding principle behind the union catalog is that the CDP can take anything from the projects. The CDP specifies the fields that it must have in a record, but the individual projects can decide what format to supply the data in. This approach is to ensure that the collaboration is as inclusive as possible and that one large project is not prescribing the format for others. The categories are broad, using Dewey Classification.

The target audience is broadly defined by five categories: general user/visitor; K-12 lifelong learner; vocations/hobbyist; scholar/researcher; and business community. All the aspects of the CDP derive from these

categories to ensure that each has the tools and resources needed for their specific purposes. There was no formal evaluation of the target audience, but the steering group established the categories at the outset of the project and these have driven the project.

They have used both internal and external advice on project management. The CDP is set up as a not-for-profit organization. The four major institution types provide a variety of steering groups to decide on the guidelines for the participating projects. A primary board oversees the running of the CDP in general, while there are groups on collections, metadata, scan centers, website, heritage, audio standards and educational use.

The CDP uses electronic means to disseminate information, such as listservs, to the projects but has found these very quiet and therefore not the best way to encourage communication.

The CDP was established with a LSTA grant of $70,000 to create a website, declare the standards and run a pilot project. The pilot was carried out by the University of Denver, involving teachers over one week using images on the topic of labor in south Colorado coalfields. The images came from the DPL, Colorado Historical Society and faculty and helped the teachers create lesson plans around the materials. From this pilot project the CDP has adapted and adopted standards and practices.

The digitization is carried out by the different projects in different ways. They use in-house facilities as well as out-sourcing. Material is digitized in-house when it is fragile or oversized. The CDP has established scan centers in the state for the use of those projects which lack the technology in-house. They have flatbed scanners, slightly larger than A4. The project trains in the use of the scanning and the use of the centers, while each institution has to do their own scanning. The scan centers are located in central points in the state allowing access for all the geographically dispersed projects. The workshops run by the CDP explain how to calibrate and how to use benchmarks in the scanning process.

8.3 Human Resources and Training

There are only two full time members of staff in the CDP, the Director and the Operations co-ordinator, and between them they cover the roles of curator, digitizer and technical development. They rely on the work of the committees to assist them with strategy and decisions on guidelines. Both have cross cultural heritage experience, which is invaluable when dealing with the variety of sub-projects. They have assistance from the Colorado Alliance of Research Libraries and the Colorado State Library for technical development, such as the union catalog and the website.

Training for these staff and for staff in the projects is on-going; full time staff receive training both on external courses and in house. Training topics include: project management; application of technical standards; preparation and handling of materials; technical operation of digitizing equipment; metadata creation; and digital preservation.

8.4 Project Life-Cycle Processes and Procedures

The copyright issue is dealt with on an individual project basis and the digital deliverables available to users vary from project to project.

The CDP has addressed the issue of preservation by using a panel of experts (one of the committees) to look at the existing resources and how each project preserves the digital material.

The catalogs for each project will vary in depth and format depending on the type and size of institution. These can range from an uncataloged box of 4,000 photographs to a full inventory and catalog of a collection.

The CDP recommends that the projects catalog the digital object under the source object. The metadata for the digital object is based on guidelines from Dublin Core and MARC. However, the projects can decide their own level of metadata, as long as they adhere to the CDP guidelines. There are 18 CDP Dublin Core elements recommended to the projects, but these can be delivered in any format to the union catalog. These elements cover the digital object, digitization process, technical details, staffing details, and administrative details. It will vary from project to project as to who creates the metadata.

The relationship between the digital objects and catalog varies with each project, however copies of metadata for each digital object or website section is sent to the CDP for it’s union catalog. All images are locally stored

and serviced. The project also allows institutions that have a separate Z39.50 accessible database of digital objects and metadata to be searched at the same time as the CDP union catalog database.

8.5 Format, Resolution and Compression of Digitized Materials

The CDP has basic guidelines on format, resolution and compression, and different projects use different levels for the various materials. Examples include: PDF for text as images; TIFF for capture and preservation; GIF, JPEG and PDF for delivery of images. The minimum resolution for images is 300dpi for capture, 75dpi delivery of thumbnails and 75-150dpi for larger delivery items. Bit depth varies between 1, 8 and 24. Both compression and delivery are via GIF or JPEG. Compression is used to improve access, decrease storage and to reduce theft. They do not recommend that the projects carry out post imaging processing because of its huge time implications. File sizes vary from project to project.

At present no projects use audio or video but they are working on guidelines for these. They are also developing guidelines for quality control and will be offering training on this from a project management perspective.

Users have access to the catalog and materials through the union catalog, but access to the master images is controlled by the participating institution. The union catalog uses Z39.50 technology to implement this access. The CDP is considering e-commerce for some projects where this is viable and it is possible to purchase the high-resolution images.

8.6 Evaluation, Funding and Long-term Sustainability

Front-end evaluation for the CDP took the form of the pilot study and the focus on the five user groups. This has adjusted the project direction and led to more specific purposes and goals. Training sessions also offer opportunities for outcome based evaluation and the CDP responds to the project needs through these outcomes. There are also focus group discussions in the various constituencies.

Since 1999 the CDP has received a variety of funding for the organizational project as well as for the project partners, totaling almost a $1 million – $177, 000 from the Colorado State Library and $550, 000 from the Institute of Museums and Library Services (IMLS) and $36,000 from the Regional Library Systems of Colorado. More funding would enable CDP to employ one more person to deal with administrative issues as well as more technical help on the website. Financial and progress reports are produced for IMLS every six months. The projects have to complete a grant form at the onset of a sub project, with a detailed budget plan and a quarterly report to the CDP.

The CDP recognizes that long term sustainability is an area to be explored. At the start it recommended storage on CDs, now it asks projects to send the digital objects via FTP to the CDP for storage on a live server as well as on CD. This is for back up as well as to insure against disaster and for recovery.

The Project is now developing a plan for operation systems migration. IMLS and NEH wish to ensure that the digital objects remain accessible in the long term due to the unique nature of the resources and so are the main influence behind this project. This sustainability depends on self-generating funds and they are considering a membership based organization to generate revenue. They still have a year of full funding remaining so this strategy is not tested yet. The possibility of an exit strategy is still under discussion if it were required, possibly moving the management to the state Library or to the distributed projects. As with all projects that involve commitment and resources, the loss of the digital deliverables would be a matter of concern.

[5] http://coloradodigital.coalliance.org/intro.html#Project

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

1. Project Management

Capture your collection: a guide for managers planning and implementing digitization projects. Ottawa: Canadian Heritage Information Network, 2000.

“Collection Management Issues and the Cedars Project.” *Based on a brainstorming meeting held in London on the 16th April, 1999,* CEDARS, www.leeds.ac.uk/cedars/documents/CIS04.htm, (acc. Oct 2000), (created April 1999).

“Conservation Implications of Digitization Projects.” 1999, Library of Congress, memory.loc.gov/ammem/techdocs/conserv83199a.pdf, (acc. Oct 2000), (created Sept 1999).

“Decision Matrix/Workflows.” Oxford University, www.bodley.ox.ac.uk/scoping/matrix.htm, (acc. Oct 2000).

“Digitisation: A Project Planning Checklist.” Arts and Humanities Data Service (AHDS), ahds.ac.uk/resource/checklist.html, (acc. Oct 2000), (created Nov 1998).

“Higher Education Digitisation Service (HEDS).” heds.herts.ac.uk/, (acc. Oct 2000).

“Planning Digital Projects for Historical Collections in New York State.” New York Public Library, digital.nypl.org/brochure, (acc. Oct 2000), (created Aug 1999).

“Porträtsammlung, Bildarchiv und Fideikommißbibliothek.” The National Library of Austria, www.onb.ac.at/sammlgn/sapofrs.htm, (acc. Oct 2000).

“Preservation Resources Scanning Questionnaire.” Online Computer Library Center, Preservation Resources, www.oclc.org/presres/scanning/scanquestion.htm, (acc. Oct 2000).

“Preservation Resources, Scanning Questionnaire.” *Quality Control Documentation,* Online Computer Library Center, Preservation Resources, www.oclc.org/presres/scanning/qa.htm, (acc. Aug 2000).

“Questions to Consider Before Beginning an Image Database Project.” University of Illinois Library, Digital Imaging & Media Technology Initiative, images.library.uiuc.edu/resources/20questns.html, (acc. Oct 2000), (created Dec 1999).

“RLG Guidelines for Creating a Request Proposal for Digital Imaging.” Research Libraries Group, www.rlg.org/preserv/RFPGuidelines.pdf, (acc. Oct 2000), (created May 1998).

“RLG Model Request for Proposal (RFP) for Digital Imaging Services.” Research Libraries Group, www.rlg.org/preserv/RLGModelRFP.pdf, (acc. Oct 2000), (created July 1998).

“RLG Tools for Digital Imaging.” Research Libraries Group, www.rlg.org/preserv/RLGtools.html, (acc. Oct 2000), (created June 1998*).

“RLG Working Group on Preservation Issues of Metadata: Final Report.” Research Libraries Group, www.rlg.org/preserv/presmeta.html, (acc. Oct 2000), (created May 1998).

"Selection Criteria, Guidelines, Decision-making Aids." Research Libraries Group, www.rlg.org/preserv/joint/selection.html, (acc. Oct 2000), (created Sept 1998).

"Technical Recommendation for Digital Imaging Projects." *Image Quality Working Group of ArchivesCom, a joint Libraries/AcIS Committee,* 1997, University of Columbia, www.columbia.edu/acis/dl/imagespec.html, (acc. Oct 2000), (created April 1997*).

"Technical Review: Outsourcing Film Scanning and Computer Output Microfilm (COM) Recording." *RLG DigiNews* 1, no. 2, August (1997). www.rlg.org/preserv/diginews/diginews2.html#hardware&software, (acc. Oct 2000).

"The Zoom Project." University of California at Berkeley, now.cs.berkeley.edu/Td/zoom.html, (acc. Oct 2000).

"Workflow Guidelines: 2. Pre-scanning Preparation and Post-scanning Processing of the Originals." TASI, www.tasi.ac.uk/framework/collections/jidi-workflow.html#wg2, (acc. Oct 2000).

Adcock, E. P., (ed.) *IFLA Principles for the Care and Handling of Library Material*. 1998, Paris/Washington.

Alexander, B., A. Craig, B. Sandore, and D. VandeCreek. *First steps in digitizing: questions to consider before beginning your project*. Springfield Ill: Illinois State Library, 1998. Video cassette.

Arcoli, A., and B. Washburn. "Delivering the AMICO Library® through RLGs Eureka® Service." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 128-129.

Armitage, L. H., and P. G. B. Enser. "Analysis of User Need in Image Archives." *Journal of Information Science* 23, no. 4 (1997): 287-299.

Arms, C. "Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress." *D-Lib Magazine* 2, no. 5, May (1996). www.dlib.org/dlib/may96/loc/05c-arms.html, (acc. Oct 2000).

Arms, W. Y. "Implementing Policies for Access Management." *D-Lib Magazine* 4, no. 2, February (1998). www.dlib.org/dlib/february98/arms/02arms.html, (acc. Oct 2000).

Arms, W. Y. "Key Concepts in the Architecture of the Digital Library." *D-Lib Magazine* 1, no. 7, July/August (1995). www.dlib.org/dlib/July95/07arms.html, (acc. Oct 2000).

Arms, W. Y., C. Blanchi, and E. Overly. "An Architecture for Information in Digital Libraries." *D-Lib Magazine* 3, no. 2, February (1997). www.dlib.org/dlib/february97/cnri/02arms1.html, (acc. Oct 2000).

Association for Information and Image Management. *Practice for Operational Procedures/Inspection and Quality Control of First-Generation, Silver Microfilm of Documents.* Silver Spring, MD: Association for Information and Image Management, 1998.

Atkinson, R. "Selection for Digitizing in Academic Libraries: One Way to Go." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 21.

Atkinson, R. "Selection for Preservation: A Materialistic Approach." *Library Resources and Technical Services* 30 (1986): 34-62.

Ayris, P. "Guidance for Selecting Materials for Digitisation." In *Presented at the Joint RLG and NPO Preservation Conference, Guidelines for Digital Imaging, September 28-30*, RLG, 1998.

Ayris, P. "Guidance for selecting materials for digitisation." In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 20-27.

Barber, D. "OhioLINK: A Consortial Approach to Digital Library Management." *D-Lib Magazine* 3, no. 4, April (1997). www.dlib.org/dlib/april97/04barber.html, (acc. Oct 2000).

Beagrie, N., and D. Greenstein. "A Strategic Policy Framework for Creating and Preserving Digital Collections." 1998, Arts and Humanities Data Service, www.ahds.ac.uk/manage/framework.htm, (acc. Jan 2000).

Beagrie, N., and D. Greenstein. "Managing Digital Collections: AHDS Policies, Standards and Practices Version

1.” 1998, Arts and Humanities Data Service, www.ahds.ac.uk/public/srg.html, (acc. Oct 2000), (created Dec 1998).

Bearman, D. “NISO/CLIR/RLG Technical Metadata Elements for Images Workshop, April 18-19, 1999.” NISO (National Information Standards Organization), www.niso.org/imagerpt.html, (acc. Oct 2000), (created April 1999).

Beckwith, B. R. “Issues Involved in Collections Management and Digitization of Images: a Review of the Literature.” Drexel University, www.pages.drexel.edu/undergrad/brb24/rol.htm, (acc. Oct 2000), (created Feb 1999).

Bellinger, M. “Outsourcing Digitization: A Service Provider’s Perspective.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 158-159.

Berger-Lefebure, R. “Bulk Conversion: Systems Tools and Experiences.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 11-12.

Bescós, J., and J. Navarro. “La Digitalización como Medio para la Preservación y Acceso a la Información en Archivos y Bibliotecas.” *Educación y Bibliotecas* 80 (1987): 28-41.

Besser, H. “Adding an Image Database to an Existing Library and Computer Environment: Design and Technical Considerations.” In *ASIS 1991 Midyear Proceedings: Multimedia Information Systems 1992*, 1992.

Besser, H. “Best Practices for Image Capture.” 1999, California Digital Library, sims.Berkeley.edu/~howard/moa2/bp90.doc, (acc. Oct 2000), (created July 1999).

Besser, H. “Digital Image Distribution: A Study of Costs and Uses.” *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/10besser.html, (acc. .

Besser, H. “Procedures and Practices for Scanning.” 1998, University of California at Berkeley, sunsite.Berkeley.edu/Imaging/Databases/Scanning/, (acc. Oct 2000).

Besser, H., and R. Yamashita. “The Cost of Digital Image Distribution: The Social and Economic Implications of the Production, Distribution, and Usage of Image Data, Final Report.” 1998, University of California, Berkeley, sunsite.Berkeley.edu/Imaging/Databases/1998mellon, (acc. Oct 2000).

Bicknese, D. “Measuring the Accuracy of the OCR in the Making of America.” *(a report prepared in fulfillment of directed field experience requirements, University of Michigan, School of Information, winter 1998),* University of Michigan, moa.umdl.umich.edu/moaocr.html, (acc. Oct 2000), (created 1998).

Bikson, T. K. “Managing Digital Documents: Technology Challenges and Institutional Responses.” In *Proceedings of the International Conference of the Round Table on Archives*, Stockholm, Sweden: International Council on Archives, *forthcoming*.

Birdsey, C. L. “Creating Accessible Digital Imagery.” *D-Lib Magazine* 6, no. 4, April (2000). www.dlib.org/dlib/april00/birdsey/04birdsey.html, (acc. Oct 2000).

Bishop, A. P. “Working Towards an Understanding of Digital Library Use: A Report on the User Research Efforts of the NSF/ARPA/NASA DLI Projects.” *D-Lib Magazine* 1, no. 10, October (1995). www.dlib.org/dlib/october95/10bishop.html, (acc. Oct 2000).

Bot, M., J. Burgemeester, and H. Roes. “The Cost of Publishing an Electronic Journal: A general model and a case study.” *D-Lib Magazine* 4, no. 11, November (1998). www.dlib.org/dlib/november98/11roes.html, (acc. Oct 2000).

Bouché, N. *Digitization for scholarly use: The Boswell Papers Project at the Beinecke Rare Book and Manuscript Library.* Washington, DC: Council on Library and Information Resources, 1999.

Bourguignon, J.-P. “User-landscape: Needs and visions of users.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 7-10.

Bradley, K. *Prince: a Practical Handbook.* Oxford 1993.

Bramich, K., and J. Cannon. *Capturing the Big Picture: Strategies for Image Libraries at the National Museum of Australia.* abridged version ed.: National Museum of Australia, 1998. www.nma.gov.au/aboutus/products/bigpic.pdf, (acc. Oct 2000).

Budd, J., and B. Harloe. “Collection Development and Scholarly Communication in the 21st Century: From Collection Management to Content Management.” In *Collection Management for the 21st Century*, edited by G E Gorman and Ruth Miller, Westport, CT: Greenwood Press, 1997. 3-25.

California Heritage Collection. “Digitizing the Collection: Overview.” University of California, Berkeley, The Bancroft Library, sunsite.berkeley.edu/CalHeritage/digital.html, (acc. Oct 2000), (created Nov 1997*).

Central Computer and Telecommunications Agency (CCTA). *PRINCE2, Project Management For Business.* London: The Stationary Office, 1997.

Chadwick, J., J. H. Falk, and B. O’Ryan. “Assessing Institutional Web Sites (report prepared for the Council on Library and Information Resources, September 30, 1999).” *Collections, Content, and the Web,* 1999, Council of Library and Information Resources, http://www.clir.org/cpa/reports/pub88/appendix2.html, (acc. Oct 2000), (created Jan 2000).

Colet, L. S. *Planning an Imaging Project.* Mountain View, CA: Research Libraries Group, 2000. http://www.rlg.org/visguides/visguide1.html, (acc. Oct 2000).

Collection Policy Working Group. “Collection Policy, June 1999.” *Colorado Digitization Project,* 1999, coloradodigital.coalliance.org/users.html, (acc. Oct 2000), (created Jan 1999).

Collections, Content, and the Web. Washington, DC: Council on Library and Information Resources, 2000. www.clir.org/cpa/reports/pub88/contents.html, (acc. Oct 2000).

Colorado Digitization Project. “Digital Toolbox.” coloradodigital.coalliance.org/toolbox.html, (acc. Oct 2000), (created Dec 1999).

Colorado Digitization Project. “Market Segments and Their Information Needs.” coloradodigital.coalliance.org/users.html, (acc. Dec 2000), (created Jan 1999).

Columbia University Digital Library Collections. “Digital Imaging Projects.” Columbia University, www.columbia.edu/cu/libraries/digital/criteria.html, (acc. Oct 2000), (created Aug 2000*).

Computer Science and Telecommunications Board. “Design and Evaluation: A Review of the State-of-the-Art.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/nrc/07nrc.html, (acc. Oct 2000).

Conway, P. “Facts and Frameworks: An Approach to Studying the Users of Archives.” *American Archivist* 49 (Fall) (1986): 393-407.

Conway, P. “Production Tracking.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 160-161.

Cox, J., and M. Taleb. “Images on the Internet: Enhanced User Access.” *Database* 17, no. 4, August (1994): 18-22, 24-26.

Davis, B. “The Digital Museum and Computer-Generated Collections and Catalogs of Art.” *Aperture* 136, no. Summer (1994): 68-70.

Day, M. “Extending Metadata for Digital Preservation.” *Ariadne* 9, no. May (1997). www.ariadne.ac.uk/issue9/metadata/intro.html, (acc. Oct 2000).

De Stefano, P. “Selection for Digital Conversion.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 11-23.

Demas, S. “What Will Collection Development Do?” In *Selecting Library and Archive Collections for Digital Reformatting: Proceedings from an RLG Symposium held November 5-6, 1995, in Washington, DC*, edited by Ricky Erway, Mountain View, CA: RLG, 1996. 18.

Donovan, K. “Anatomy of an Imaging Project: A Primer for Museums, Libraries, Archives and Other Visual Collections.” *Spectra, A Publication of the Museum Computer Network* 23, no. 2, Winter (1995-6): 19-22.

Edelson, D. C., and D. N. Gordin. “Adapting Digital Libraries for Learners: Accessibility vs. Availability.” *D-Lib Magazine* 2, no. 9, September (1996). www.dlib.org/dlib/september96/nwu/09edelson.html, (acc. Oct 2000).

Endres, A. “Digital library and electronic publishing activities of German Professional societies.” In *Proceedings of the 1997 IEEE Forum on Research and Technological Advances in Digital Libraries (May 7-9, 1997, Washington DC)*, 1997. 63-68.

Federal Task Force on Digitization (Canada). *Towards a learning nation: The digital contribution: recommendations proposed by the Federal Task Force on Digitization: final report.* Ottawa: National Library of Canada, 1999.

Ferguson, I. A., and M. J. Wooldridge. “Paying Their Way: Commercial Digital Libraries for the 21st Century.” *D-Lib Magazine* 3, no. 6, June (1997). www.dlib.org/dlib/june97/zuno/06ferguson.html, (acc. Oct 2000).

Fleischhauer, C. “Selecting Collections and Selecting Technology: American Memory at the Library of Congress.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 20.

Gay, G., R. Rieger, and A. Sturgill. “Findings of the MESL Casual User Survey.” In *Delivering Digital Images: Cultural Heritage Resources for Education, The Museum Educational Site Licensing Project*, edited by Christie Stephenson and Patricia McClung, Los Angeles: The Getty Information Institute, 1998.

Gertz, J. “Guidelines for Digital Imaging.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 11-20.

Getz, M. “Questionnaire for Digital Scriptorium Users.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSquery.pdf, (acc. Oct 2000), (created March 1998).

Gillings, M., A. Wise, and (eds). “GIS Guide to Good Practice.” 1998, Arts and Humanities Data Service, London, ads.ahds.ac.uk/project/goodguides/gis/index.html, (acc. Oct 2000).

Gladney, H. M. “Safeguarding Digital Library Contents and Users: A Note on Universal Unique Identifiers.” *D-Lib Magazine* 4, no. 4, April (1998). www.dlib.org/dlib/april98/04gladney.html, (acc. Oct 2000).

Gladney, H. M. “Safeguarding Digital Library Contents and Users: Document Access Control.” *D-Lib Magazine* 3, no. 6, June (1997). www.dlib.org/dlib/june97/ibm/06gladney.html, (acc. Oct 2000).

Gladney, H. M. “Safeguarding Digital Library Contents and Users: Interim Retrospect and Prospects.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/gladney/07gladney.html, (acc. Oct 2000).

Gladney, H. M., and J. B. Lotspiech. “Safeguarding Digital Library Contents and Users: Assuring Convenient Security and Data Quality.” *D-Lib Magazine* 3, no. 5, May (1997). www.dlib.org/dlib/may97/ibm/05gladney.html, (acc. Oct 2000).

Gladney, H. M., and J. B. Lotspiech. “Safeguarding Digital Library Contents and Users: Storing, Sending, Showing, and Honoring Usage Terms and Conditions.” *D-Lib Magazine* 4, no. 5, May (1998). www.dlib.org/dlib/may98/gladney/05gladney.html, (acc. Oct 2000).

Guthrie, K. M. “JSTOR: The Development of a Cost-Driven, Value Based Pricing Model.” *(paper presented at ARL’s Scholarly Communication and Technology Conference, Emory University, April 24-25, 1997),* 1997, Association of Research Libraries, arl.cni.org/scomm/scat/guthries.html, (acc. Oct 2000), (created April 1997).

Häkli, E. “A European model: Organisation.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 85-90.

Harvard University Libraries, P. “Selection for Digitization: A Decision Making Matrix.” Harvard University, preserve.harvard.edu/bibliographies/selection.html, (acc. Oct 2000), (created March 2000).

Hastings, S. K. “Evaluation of Image Retrieval Systems: The Role of User Feedback.” *Library Trends* 48, no. Fall (1999).

Hazen, D., J. Horrell, and J. Merrill-Oldham. *Selecting Research Collections for Digitization.* Washington, DC:

Council on the Library and Information Resources, 1998. www.clir.org/pubs/reports/hazen/pub74.html, (acc. Oct 2000).

Hendley, T. *Comparison of Methods & Costs of Digital Preservation: A Consultancy Study.*, *British Library Research and Innovation Report 106.* London: British Library Research and Innovation Centre, 1998.

Hert, C. A., and G. Marchionini. "Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations." 1987, University of North Carolina, http://ils.unc.edu/~march/blsreport/mainbls.html, (acc. Oct 2000), (created July 1997).

Holm, J. "Factors to Consider in Pictorial Digital Image Processing." In *Proceedings, IS&T 49th Annual Conference*, Minneapolis, MN, 1996. 298-304.

Holm, J. "The Evaluation of Digital Photography Systems, Short Course Notes." In *IS&T 49th Annual Conference*, 1996.

Hornsby, E. L., J. Arnott, D. B. Marcum, A. R. Kenney, H. T. Hickerson, C. D. Spornick, N. Smith, J. Rader, and S. Nyberg. *To scan or not to scan: what are the questions? Proceedings of a SOLINET conference on digitizing projects for libraries and archives: May 2, 1996, Atlanta, Georgia.* Atlanta, GA: Southeastern Library Network, 1997.

Hurley, B. J., J. Price-Wilkin, M. Proffitt, and H. Besser. *The Making of America II Testbed Project: A Digital Library Service Model.* Washington, DC: The Digital Library Federation, 1999.

InforSurf: The USCB Library. "University of California Selection Criteria for Digitization." University of Southern California, library.usc.edu/ucpag/digselec.html, (acc. 2000).

Ingram, G. B., and C. Rickerson. "Selection of Special Collection Materials for Digitization at the University of Washington Libraries." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 18-19.

Jephcott, S. "Why Digitise? Principles in Planning and Managing a Successful Digitisation Project." *The New Review of Academic Librarianship* 4 (1998): 39-53.

Jones, M. L. W., G. K. Gay, and R. H. Rieger. "Project Soup: Comparing Evaluations of Digital Collection Efforts." *D-Lib Magazine* 5, no. 11, November (1999). www.dlib.org/dlib/november99/11jones.html, (acc. Oct 2000).

Jones-Garmil, K. (ed)., *The Wired Museum: Emerging Technology and Changing Paradigms.* American Association of Museums, 1997.

Kahn, R., and R. Wilensky. "A Framework for Distributed Object Services." 1995, The Corporation for National Research Initiatives (CNRI), www.cnri.reston.va.us/home/cstr/arch/k-w.html, (acc. Oct 2000), (created May 1995).

Kenney, A. R. "Conversion of Traditional Source Materials into Digital Form." In *Research Agenda for Networked Cultural Heritage*, edited by David Bearman, Santa Monica, CA: Getty Art History Information Program, 1996. 41-47.

Kenney, A. R. "Digital Benchmarking for Conversion and Access." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 24-60.

Kenney, A. R. "Guidelines vs. Guidance for Digital Imaging: The Opportunity Before Us." In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 1-10.

Kenney, A. R. "Projects to Programs: Mainstreaming Digital Imaging Initiatives." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 153-175.

Kenney, A. R., and L. H. Sharpe, II. "Illustrated Book Study: Digital Conversion Requirements of Printed Illustrations." 1999, Library of Congress, Preservation, www.lcweb.loc.gov/preserv/rt/illbk/ibs.htm, (acc. Oct 2000), (created July 1999).

Kenney, A. R., and O. Y. Rieger. *Managing Digital Imaging Projects: An RLG Workshop.*: RLG, 1998.

Kessler, B. R. “Electronic Images in Visual Resources Collection: Some Strategic Question.” *Visual Resources* X (1994): 1-10.

Kilker, J., and G. Gay. “The Social Construction of a Digital Library: A Case Study Examining Implications for Evaluation.” *Information Technology and Libraries* 17, no. 2 (1998): 60-69.

Kingma, B. R. “The Costs of Print, Fiche, and Digital Access: The Early Canadiana Online Project.” *D-Lib Magazine* 6, no. 2, February (2000). www.dlib.org/dlib/february00/kingma/02kingma.html, (acc. Oct 2000).

Klijn, E., and Y. de Lusenet. *In the Picture: Preservation and Digitisation of European Photographic Collections.* Amsterdam: European Commission on Preservation and Access, 2000.

Knoll, A. “For Whom and How We Digitize.” *Web World, Memory of the World, Memoriae Mundi Series, Bohemica,* 1998, UNESCO, www.unesco.org/webworld/mdm/czech_digitization/doc/users.htm#ForWhom%20To, (acc. Oct 2000).

Lagoze, C., and D. J. Dienst. “An architecture for distributed document libraries.” *Communications of the ACM* 38, no. 4, April (1995): 47.

Lee, S. D. “Assessment Criteria for Digitization.” 1999, Oxford University, www.bodley.ox.ac.uk/scoping/assessment.html, (acc. Oct 2000), (created Jan 1999).

Lee, S. D. “Scoping the Future of The University of Oxford’s Digital Library Collections.” 1999, Oxford University, www.bodley.ox.ac.uk/scoping/report.html, (acc. Oct 2000), (created Sept 1999).

Levy, D. M. “What Do You See and What Do You Get? Document Identity and Electronic Media.” In *Screening Words: User Interfaces for Text; Proceedings of the Eighth Annual Conference of the UV Centre for the New OED and Text Research*, 1992. 109-117.

Library of Congress, and Contracts and Logistics Services. “The Library of Congress Requests Proposals for Digital Images of Pictorial Materials, Solicitation No. RFP97-9.” memory.loc.gov/ammem/prpsal9/coverpag.html, (acc. Oct 2000).

Library of Congress: National Digital Library Program and the Conservation Division. “Conservation Implications of Digitization Projects.” Library of Congress, American Memory, memory.loc.gov/ammem/techdocs/conservation.html, (acc. Oct 2000), (created Sept 1999).

Library of Congress: National Digital Library Program. “NDLP Project Planning Checklist.” Library of Congress, American Memory, memory.loc.goc/ammem/prjplan.html, (acc. Oct 2000), (created Jan 1997*).

Lunin, L. F. “The Big Picture: Selection and Design Issues for Image Information Systems.” In *Digital Image Access & Retrieval*, edited by P Bryan Heidorn and Beth Sandore, Urbana, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1997.

Lynch, C. “Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information.” *D-Lib Magazine* 5, no. 9, September (1999). www.dlib.org/dlib/september99/09lynch.html, (acc. Oct 2000).

MacKie-Mason, J. K., M. S. Bonn, J. F. Riveros, and W. P. Lougee. “A Report on the PEAK Experiment: Usage and Economic Behaviour.” *D-Lib Magazine* 5, no. 7/8 (1999). www.dlib.org/dlib/july99/mackie-mason/07mackie-mason.html, (acc. Oct 2000).

Macklin, L. L., and S. L. Lockmiller. *Digital Imaging of Photographs. A Practical Approach to Workflow Design and Project Management.*, *LITA Guides no. 4.* Chicago: American Library Association, 1999.

Maly, K., M. L. Nelson, and M. Zubair. “Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework.” *D-Lib Magazine* 5, no. 3, March (1999). www.dlib.org/dlib/march99/maly/03maly.html, (acc. Oct 2000).

Marchionini, G., C. Plaisant, and A. Komlodi. “Interfaces and Tools for the Library of Congress National Digital Library Program.” *Information Processing and Management* 34, no. 5 (1998): 535-555.

McIntyre, J. E. “Protecting the Physical Form.” In *Guidelines for Digital Imaging: papers given at the Joint*

National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998, London: The National Preservation Office, 1998. 33-38.

McKnight, C., and A. Dillon. “User-centred design of library information systems: HyperLib.” In *Van Geautomatiseerd Beheer van Archieven en Bibliotheken naar Geautomatiseerde Informatie*, edited by K Clara and J van Borm, Antwerp: Vlaamse Vereniging voor Bibliotheek-, Archief- an Documentatiewezen, 1993. 163-170.

Morris, R. J. “Project Design and the Creation of Digital Records in Scotland.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 150-163.

Orna, E., and Pettitt, C. *Information Management in Museums.* 2nd edition, Gower Publishing Company, 1998.

Pitti, D. V. “Access to Digital Representations of Archival Materials: The Berkeley Finding Aid Project.” In *RLG Digital Image Access Project, Proceedings from an RLG Symposium held March 31 and April 1, 1995*, Palo Alto, CA, 1993. 73-81.

Popkin, J., and A. Cushman. *Integrated Document Management - Controlling a Rising Standard.* Stamford, Conn: Gartner Group, 1993.

Price-Wilkin, J. “Moving the Digital Library from ‘Project’ to ‘Production’.” *(paper presented at DLW99 in Tsukuba, Japan, February 1999),* University of Michigan, jpw.umdl.umich.edu/pubs/japan-1999.html, (acc. Oct 2000), (created Feb 1999).

Puglia, S. “The Costs of Digital Imaging Projects.” *RLG DigiNews* 3, no. 5, October 15 (1999). www.rlg.org/preserv/diginews/diginews3-5.html#feature, (acc. Oct 2000).

Puglia, S., and B. Roginski. “NARA Guidelines for Digitizing Archival Materials for Electronic Access.” 1998, National Archives and Records Administration, www.nara.gov/nara/vision/eap/digguide.pdf, (acc. Oct 2000), (created Jan 1998).

Reid, G. “The digitisation of heritage material: arguing for an interpretative approach based on the experience of the Powys Digital History Project.” *Program - Electronic Library and Information Systems* 34, no. 2, April (2000): 143-158.

Riser, J. “Hand-crafted cherry book cradles.” University of Virginia, www.lib.virginia.edu/speccol/scdc/cradle.html, (acc. Oct 2000).

Ross, S. “Funding Information and Communication Technology in the Heritage Sector: policy recommendations to the Heritage Lottery Fund.” 1998, University of Glasgow, Humanities Advanced Technology and Information Institute, www.hatii.arts.gla.ac.uk/Projects/HLF/, (acc. Oct 2000), (created Feb 1998*).

Ross, S. “Strategies for Selecting Resources for Digitization: Source-Orientated, User-Driven, Asset-Aware Model (SOUDAAM).” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 5-27.

Royan, B. “Project Management and Quality Assurance.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 28-37.

Sander, M., and J. Perkins. *Integrated Information Management Model,* CIMI Publication, June 1999. www.cimi.org/publications.html#iim_1, (acc. Oct 2000).

Schutzer, D. “A Need For A Common Infrastructure: Digital Libraries and Electronic Commerce.” *D-Lib Magazine* 2, no. 4, April (1996). www.dlib.org/dlib/april96/04schutzer.html, (acc. Oct 2000).

Seaman, D. “Goals and Missions.” University of Virginia Library, Electronic Text Center, etext.lib.virginia.edu/mission.html, (acc. Oct 2000).

Secor, J. R. “Digitizing as Strategy.” In *Economics of Digital Information, Collection, Storage, and Delivery*, edited by Sul H Lee, Binghampton, NY: The Haworth Press, 1997.

Selection Task Force of the Electronic Preservation Committee. “Selection Criteria and Guidelines: Selecting Materials for Digital Preservation at the National Agricultural Library.” 1995, National Agricultural Library, www.preserve.nal.usda.gov:8300/projects/criteria.htm, (acc. Oct 2000), (created Nov 1995).

Stokes, J. R. “Sheet Music Scanning Procedures.” Historic American Sheet Music, Duke University, Rare Book, Manuscript and Special Collections Library, http://scriptorium.lib.duke.edu/sheetmusic/procedure.html, (acc. Oct 2000).

Süsstrunk, S. “Image Production Systems at Corbis Corporation.” *RLG DigiNews* 2, no. 4, August 15 (1998). www.rlg.org/preserv/diginews/diginews2-4.html#technical, (acc. Oct 2000).

Swartzell, A. G. “Preparation of Materials for Digitisation.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 29-32.

Tanner, S. “Appendix J: Costed Collections.” In *Scoping the Future of The University of Oxford’s Digital Library Collections*, 1999.

Tanner, S. “Standards.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 9-10.

Tanner, S., and B. Robinson. “JISC Image Digitisation Initiative (JIDI) Feasibility Study (Final Report).” 1997, Higher Education Digitisation Service, Hatfield, heds.herts.ac.uk/resources/papers/jidi_fs.html, (acc. Oct 2000), (created July 1998).

Tanner, S., and J. L. Smith. “Digitisation: How Much Does it Really Cost?” *(paper for the Digital Resources for the Humanities 1999 Conference, Kings College London, September 12-15, 1999),* 1999, heds.herts.ac.uk/resources/papers/drh99.pdf, (acc. Oct 2000), (created Sept 1999).

US Army Corps of Engineers. “Economic Analysis, Executive Summary Report, Optical Disk Imaging.” United States Army Corp of Engineers, http://144.3.144.209/corpusdata/usace/inet/usace-docs/s-r/etl1110-1-168/entire.pdf, (acc. Oct 2000), (created Oct 1997).

Weise, J. “Image Database Access Architecture.” *(version I.I., draft I),* 1997, University of Michigan, images.umdl.umich.edu/info/arch/Image_Tech_Dev_Plan.html, (acc. Oct 2000), (created June 1997).

Willis, D., and D. Lathin. “Make Your Digital Imaging Program Successful: Involve the Stakeholders.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 7.

Witten, I. H., A. Moffat, and T. C. Bell. *Managing gigabytes: Compressing and Indexing Documents and Images.*: Van Nostrand Reinhold, 1994.

Woodward, H. “Selecting journals for digitisation: piecing together the puzzle to create a European model.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 77-83.

Zorich, D. M. *Introduction to Managing Digital Assets: Options for Cultural and Educational Organizations.* Los Angeles: Getty Information Institute, 1999.

content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

2. Rights Management

"ECUP (European Copyright User Platform) Copyright Focal Point." www.eblida.org/ecup, (acc. Oct 2000).

"Internet Library of Early Journals (January 1996 - August 1998). A Project in the eLib Programme: Final Report." www.bodley.ox.ac.uk/ilej/papers/fr1999/fr1999.htm, (acc. Oct 2000), (created March 1999).

"Rights Management Framework." Arts and Humanities Data Service (AHDS), ahds.ac.uk/bkgd/rmf1.htm, (acc. Oct 2000), (created Dec 1998*).

Balas, J. "Copyright in the Digital Era." *Computers in Libraries* 18, no. 6, June (1998).

Baron, R. "Copyright Law in the Digital World." 1998, College Art Association, Toronto, Canada, http://www.pipeline.com/~rabaron/ttm/TTM.htm, (acc. Oct 2000), (created Feb 1998).

Bearman, D. A. "Model Agreement for Owners of Images Licensing to Multimedia Producers." *Archives and Museum Informatics* 5, no. 2, Summer (1991): 9-10.

Berghel, H., and L. O'Gorman. "Protecting ownership rights through digital watermaking." *Computer* 7, July 29 (1996): 101-103.

Besser, H. "MESL Implementation at the Universities." In *Delivering Digital Images: Cultural Heritage Resources for Education, The Museum Educational Site Licensing Project*, edited by Christie Stephenson and Patricia McClung, Los Angeles: The Getty Information Institute, 1998. 70-85.

Braudaway, G., K. Magerlein, and F. Mintzer. "Protecting publicly available images with a visible image watermark." In *Proceedings of Optical Security and Counterfeit Deterrence Techniques*, Bellingham, WA: SPIE, 1996. 287-289.

Cohen, J. E. "Some reflections on copyright management systems and laws designed to protect them." *Berkeley Tech. Law Journal* 12, no. 1 (1997): 161-188.

Cornish, G. "Copyright, digitisation and preservation." Library Conservation News, no. 53 (1997): Winter.

Cornish, G. P., (ed.) *Copyright: Interpreting the law for libraries, archives and information services*. second ed. 1997: Library Association Publishing.

Demas, S. G., and J. L. Brogdon. "Determining Copyright Status for Preservation and Access: Defining Reasonable Effort." *Library Resources & Technical Services* 41, no. 4, October (1998): 323-334.

Digitisation and the challenge to copyright: CLIP seminar Thursday 28 April, 1994, *CLIP seminar report*. 1994: Intellectual Property Institute.

Ginsburg, J. C. "Copyright Without Walls?: Speculation on Literary Property in the Library of the Future." *Representations Special Issue: Future Libraries* 42 (1993): 53-67.

Gladney, H. M. “Digital Dilemma: Intellectual Property. Synopsis and Views on the Study by the National Academies’ Committee on Intellectual Property Rights and the Emerging Information Infrastructure.” *D-Lib Magazine* 5, no. 12, December (1999). www.dlib.org/dlib/december99/12gladney.html, (acc. Oct 2000).

Gladney, H. M., and J. B. Lotspiech. “Safeguarding Digital Library Contents and Users: Storing, Sending, Showing, and Honoring Usage Terms and Conditions.” *D-Lib Magazine* 4, no. 5, May (1998). www.dlib.org/dlib/may98/gladney/05gladney.html, (acc. Oct 2000).

Gladney, H. M., F. Mintzer, and F. Schiattarella. “Safeguarding Digital Library Contents and Users. Digital Images of Treasured Antiquities.” *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/vatican/07gladney.html, (acc. Oct 2000).

Harper, G. “What About Copyright?” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 12-13.

HMSO. “Chapter 48.” In *Copyright Designs and Patents Act, 1988*: HMSO, 1988.

Johnson, N. F., and S. Jajodia. “Exploring Steganography: Seeing the Unseen.” *Computer* February 31, no. 2 (1998): 26-34.

Kohl, U., J. Lotspiech, and M. A. Kaplan. “Safeguarding Digital Library Contents and Users: Protecting Documents Rather Than Channels.” *D-Lib Magazine* 3, no. 9, September (1997). www.dlib.org/dlib/september97/ibm/09lotspiech.html, (acc. Oct 2000).

Lu, K. “Technological Challenges to Artists’ Rights in the Age of Multimedia: The Future Role of Moral Rights.” *Reference Services Review* 22, no. 1 (1994): 9-19.

Mintzer, F., J. Lotspiech, and N. Morimoto. “Safeguarding Digital Library Contents and Users: Digital Watermarking.” *D-Lib Magazine* 3, no. 12, December (1997). www.dlib.org/dlib/december97/ibm/12lotspiech.html, (acc. Oct 2000).

Oppenheim, C., (ed.) *The Legal and Regulatory Environment for Electronic Information*. third ed. 1999: Infonortics.

Samuelson, P. “Encoding the Law into Digital Libraries.” *Communications of the ACM* 41, no. 4, April (1998): 13-18.

Stefik, M. “Shifting the possible: How trusted systems and digital property rights challenge us to rethink digital publishing.” *Berkeley Tech. Law Journal* 12, no. 1 (1997).

Steiner, C. “Controlling Your Images - The Museum and the Licensing of Imaging Products.” *Museum News* 71, no. 4, July/August (1992): 62-64.

Steiner, C., (ed.). *A Museum Guide to Copyright and Trademark.* American Association of Museums, 1999.

Steiner, C., and J. D. Neuburger. “New Media and Emerging Technologies - Staying Interactive in the High Tech Environment, a White Paper and Licensor Orientated Content License Agreement.” In *Archives and Museum Informatics. Hands on Hypermedia and Interactivity in Museums. Selected papers from the Third International Conference (ICHIM95/MCN95)*, San Diego, CA, 1995. 219-248.

Tennant, R. “Beg, Buy, Borrow, License or Steal.” *Library Journal Digital* July 15 (2000). http://www.ljdigital.com/articles/infotech/digitallibraries/20000615_15167.asp, (acc. Oct 2000).

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

3. Mark-Up (text; metadata)

"American Memory DTD for Historical Documents." Library of Congress, American Memory, memory.loc.gov/ammem/amdtd.html, (acc. Oct 2000), (created June 1998).

"Digital Page Imaging and SGML: An Introduction to the Electronic Binding DTD (Ebind)." Berkeley Digital Library SunSITE, sunsite.berkeley.edu/Ebind/index.html, (acc. Oct 2000), (created June 1996).

"Ein neues Konvertierungsprogramm der SUB Göttingen/GDZ zur Umwandlung bibliographischer Pica/GBV-Daten in RDF/XML." Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/gdz-tools/pica2xml/pica2xml_index_de.html, (acc. Oct 2000), (created 1997).

"Encoded Archival Description Official Web Site." Library of Congress, Network Development & MARC Standards Office, lcweb.loc.gov/ead, (acc. Oct 2000), (created Sept 2000*).

"Extensible Markup Language (XML)." World Wide Web Consortium, www.w3.org/XML/, (acc. Oct 2000), (created April 1997).

"Guidance on expressing the Dublin Core within the Resource Description Framework (RDF)." Dublin Core Metadata Initiative, www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/, (acc. Oct 2000), (created July 1999*).

"Hypertext Markup Language home page." World Wide Web Consortium, www.w3.org/MarkUp/, (acc. Oct 2000), (created Oct 2000*).

"JCCs SQL std. Page." JCC Consulting Inc., www.jcc.com/SQLpages/jccs_sql.htm, (acc. Oct 2000), (created May 1997).

"Overview of SGML Resources." World Wide Web Consortium, www.w3.org/MarkUp/SGML, (acc. Oct 2000), (created Nov 1995).

"Resource Description Framework (RDF)." Word Wide Web Consortium, www.w3.org/RDF/, (acc. Oct 2000).

"RLG Working Group on Preservation Issues of Metadata: Final Report." Research Libraries Group, www.rlg.org/preserv/presmeta.html, (acc. Oct 2000), (created May 1998).

"Sample Structural Metadata for the Coolidge Image-Only Documents." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/timag-samp.html, (acc. Oct 2000), (created March 1997).

"Sample Structural Metadata for the Coolidge Photos." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/photo-samp.html, (acc. Oct 2000), (created March 1997).

"Sample Structural Metadata for the Coolidge SGML Encoded Documents." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/sgml-

samp.html, (acc. Oct 2000), (created April 1997).

“SQL FAQ: Table of Contents.” *Sequoia 2000 Illustra Frequently Asked Questions,* Berkeley University of California, epoch.cs.berkeley.edu:8000/sequoia/dba/montage/FAQ/SQL_TOC.html, (acc. Oct 2000), (created June 1994*).

“TEI and XML in Digital Libraries.” *A report on a two-day workshop hosted by the Library of Congress,* University of Michigan, www.umdl.umich.edu/workshops/teidlf/sumrec.html, (acc. Oct 2000).

“TEI Text Encoding in Libraries: Draft Guidelines for Best Practices.” *version 1.0,* Indiana University, www.indiana.edu/~letrs/tei, (acc. Oct 2000), (created July 1999).

“The Data Documentation Initiative: A Project to Develop an XML Document Type Definition for Data Documentation.” ICPR, www.icpsr.umich.edu/DDI/code-book.html, (acc. Oct 2000), (created March 2000).

“What is the Best OCR? OCR Software? Scanner Survey Results.” hosc.net, www.hosc.net/ocr, (acc. Oct 2000), (created July 1997*).

“XHTML™ 1.0: The Extensible HyperText Markup Language. A Reformulation of HTML 4 in XML 1.0. W3C Recommendation 26 January 2000.” World Wide Web Consortium, www.w3.org/TR/xhtml1/, (acc. Oct 2000), (created Jan 2000).

Beech, D., S. Lawrence, M. Maloney, N. Mendelsohn, and H. S. Thompson, (eds.). “XML Schema Part 1: Structures, W3C Working Draft.” 1999, World Wide Web Consortium, http://www.w3.org/1999/05/06-xmlschema-1/, (acc. Oct 2000), (created May 1999).

Biron, P. V., A. Malhotra, (eds.). “XML Schema Part 2: Datatypes, W3C Working Draft.” 1999, World Wide Web Consortium, http://www.w3.org/1999/05/06-xmlschema-2/, (acc. Oct 2000), (created May 1999).

Bray, T., J. Paoli, and C. M. Sperberg-McQueen, (ed.). “Extensible Markup Language (XML) 1.0: W3C Recommendation 10 February 1998.” 1998, World Wide Web Consortium, www.w3.org/TR/1998/REC-xml-19980210, (acc. Oct 2000), (created Feb 1998).

Bureau of Canadian Archivists. *Planning Committee on Descriptive Standards. Subject Indexing for Archives. The Report of the Subject Indexing Working Group.* Ottawa: Bureau of Canadian Archivists, 1992.

Burnard, L., and C. M. Sperberg-McQueen. “TEI Lite: An Introduction to Text Encoding for Interchange.” *Document no. TEI U 5,* 1995, www-tei.uic.edu/orgs/tei/intros/teiu5.html, (acc. Oct 2000), (created June 1995).

Coleman, J., and D. Willis. *SGML as a Framework for Digital Preservation and Access.* Washington, DC: Commission on Preservation and Access, 1997.

Coles, T., A. Alexaner, and G. Shaw. “Following the Script: Optical Character Recognition Technology and the British Town and Trade Directory.” *History and Computing* 9 (1997): 1-16.

Dahl, K. “OCR Trends and Implications.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 111-112.

Duggan, M. K. “19th-Century California Sheet Music.” University of California, Berkeley, www.sims.berkeley.edu/~mkduggan/project.html, (acc. Oct 2000).

Getz, M. “Questionnaire for Digital Scriptorium Users.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSquery.pdf, (acc. Oct 2000), (created March 1998).

Getz, M. “The Digital Scriptorium: Medieval Manuscripts on the World Wide Web.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSvision.pdf, (acc. Oct 2000), (created March 1998).

Gill, T., C. Grout, and L. Smith. “Visual arts, museums and cultural heritage information standards: a domain-specific review of relevant standards for networked information discovery.” Visual Arts Data Service, vads.ahds.ac.uk/training_advice/standards/standards_review/info%20standards.html, (acc. Oct 2000), (created March 1997).

Gill, T., A. Gilliland-Swetland, and M. Baca. *Introduction to Metadata. Pathways to Digital Information*. Getty Institute, 1998. www.getty.edu/gri/standard/intrometadata/, (acc. Oct 2000).

Hara, S., and H. Yasunaga. *"Markup and Conversion of Japanese" D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/japan/07hara.html, (acc. Oct 2000).

Hüfner, J. "Steigerung der Erkennungsgenauigkeit durch maschinellen Abgleich verschiedener durch OCR erzeugter Volltexte (Increasing the Quality of OCR-generated Full Texts by Automatic Comparison)." *Nachrichten für Dokumentation* 48, no. 2 (1997): 79-85.

Ide, N. M., C. M. Sperberg-McQueen, and M. E. Taylor, (eds.) *TEI10, a conference in celebration of the tenth anniversary of the Text Encoding Initiative: Brown University, Providence, Rhode Island, November 14-16, 1997*. 1997, Providence, RI: Brown University.

Kirschenbaum, M. "Documenting Digital Images: Textual Meta-Data at the Blake Archive." *The Electronic Library* 16, no. 4, August (1998): 239-241.

Levy, D. M. "What Do You See and What Do You Get? Document Identity and Electronic Media." In *Screening Words: User Interfaces for Text; Proceedings of the Eighth Annual Conference of the UV Centre for the New OED and Text Research*, 1992. 109-117.

Mantelman, L. "Windows OCR Packages Put a Typist in Your Tank." *Imaging Magazine* December (1994): 8-25.

Miller, E. "An Introduction to the Resource Description Framework." *D-Lib Magazine*, no. May (1998). www.dlib.org/dlib/may98/miller/05miller.html, (acc. Oct 2000).

Miller, P. "Metadata for the Masses." *Ariadne* 5. www.ariadne.ac.uk/issue5/metadata-masses/intro.html, (acc. Oct 2000).

Morris, R. J. "Project Design and the Creation of Digital Records in Scotland." In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 150-163.

National Council on Archives. *Rules for the construction of personal, place and corporate names.*: NCA, 1997.

Oard, D. W. "Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries." *D-Lib Magazine* 3, no. 12, December (1997). www.dlib.org/dlib/december97/oard/12oard.html, (acc. Oct 2000).

Paskin, N., and G. Rust. "The Digital Object Identifier Initiative: Metadata Implications." *DOI Discussion Paper 2, version 3,* 1999, dx.doi.org/10.100/104, (acc. Oct 2000), (created Feb 1999).

Phelps, T., and R. Wilensky. "Toward active, extensible, networked documents: Multivalent architecture and applications." In *Proceedings of DL'96*, New York: ACM Press, 1996. 100-108.

Pitti, D. V. "Encoded Archival Description: An Introduction and Overview." *D-Lib Magazine* 5, no. 11, November (1999). www.dlib.org/dlib/november99/11pitti.html, (acc. Oct 2000).

Price-Wilkin, J. "Just-in-Time Conversion, Just-in-Case Collections: Effectively Leveraging Rich Document Formats for the WWW." *D-Lib Magazine* 3, no. 5, May (1997). www.dlib.org/dlib/may97/michigan/05pricewilkin.html, (acc. Oct 2000).

Rinehart, R., and A. Gilland_Swetland, (eds.). *Archives and Museum Informatics* special issue, *EAD in Action* 12, no. 3-4.

Robinson, P. *The Transcription of Primary Textual Sources Using SGML.* Oxford: Office for Humanities Communication Publications, No 6, 1994.

Rothenberg, J. "Metadata to Support Data Quality and Longevity." *Proceedings of the 1st IEEE Metadata Conference, NOAA Complex, Silver Spring, Md., 16-18 April,* 1996, www.computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html, (acc. Oct 2000).

Rust, G., and M. Bide. "The <indecs> Metadata Model." *version no. 3,* 1999, indecs, www.indecs.org/results/archive.htm, (acc. Oct 2000), (created July 1999).

Seaman, D. "Guidelines for SGML Text Mark-up at the Electronic Text Center." University of Virginia Library, Electronic Text Center, etext.lib.virginia,edu/tei/uvatei11.html, (acc. Oct 2000).

Society of American Archivists, *Encoded Archival Description Application Guidelines*. Society of American

Archivists, 2000.

Sosteric, M. “ICAAP eXtended Markup Language: Exploiting XML and Adding Value to the Journals Production Process.” *D-Lib Magazine* 5, no. 2, February (1999). www.dlib.org/dlib/february99/02sosteric.html, (acc. Oct 2000).

van Horik, R. “Recent progress in the automatic reading of printed historical documents.” In *History and Computing 5*, edited by P K Doornvan and R van Horik, 1993. 68-73.

Weibel, S. L., and C. Lagoze. “An Element Set to Support Resource Discovery: The State of the Dublin Core.” *International Journal on Digital Libraries* 1, no. 2 (1997): 176-186.

Weibel, S., and J. Hakala. “DC-5: The Helsinki Metadata Workshop: A Report on the Workshop and Subsequent Developments.” *D-Lib Magazine* 4, no. 2, February (1998). www.dlib.org/dlib/february98/02weibel.html, (acc. Oct 2000).

Witten, I. H., C. Nevill-Manning, R. McNab, and S. J. Cunningham. “A Public Library Based on Full-text Retrieval.” *Communications of the ACM* 41, no. 4, April (1998): 71-75.

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

4. Images (capture; management)

"£1m lottery grant for art database." *Library Association Record* 99, no. 11, November (1997): 584.

"About JSTOR." Journal Storage (JSTOR), www.jstor.ac.uk/about/need.html, (acc. Oct 2000), (created 2000).

"AMICO Data Specification." AMICO, www.amico.net/docs/dataspec.html, (acc. Oct 2000).

"AMICO: Art Museum Image Consortium home page." AMICO, www.amico.net, (acc. Oct 2000).

"An Introduction to Photos. 19th and Early 20th Century Photographic Processes." 1996, Public Record Office Kew, www.pro.gov.uk/preservation/photos.htm, (acc. Oct 2000), (created July 1999*).

"Bildarchiv zur Kunst und Architektur (The Image Archive for Art and Architecture)." University of Marburg, bildindex.de/intro.htm, (acc. Oct 2000).

"'Bildarvet Project', Digitalizing the Finish-Swedish heritage of pictures for research." Âbo Akademi University, www.abo.fi/library/bildarvet/welcomee.sht, (acc. Oct 2000), (created Oct 1998).

"California Digital Library: Digital Collection Image Standards." 1999, California Digital Library, www.ucop.edu/irc/cdl/tasw/Current/Imaging.Stds-090199/Imaging.Stds-090199.doc, (acc. Oct 2000), (created Sept 1999).

Capture your collection: a guide for managers planning and implementing digitization projects. Ottawa: Canadian Heritage Information Network, 2000.

"Cartesian Perceptual Progression: Technical Information." Cartesian Products, Inc., www.cartesianinc.com/Tech, (acc. Oct 2000), (created 1998).

"Center for Retrospective Digitization, Göttingen State and University Library." Göttinger Digitalizierungs-Zentrum, www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html, (acc. Oct 2000), (created June 1999).

"Centre for Photographic Conservation (UK)." dspace.dial.pipex.com/cpc.moor/, (acc. Oct 2000), (created 1999).

"CGSD Service, Gamma Literature." CGSD (Computer Graphics Systems Development Corporation), www.cgsd.com/papers/gamma.html, (acc. Oct 2000), (created 2000).

"Chapter II, Color Theory." Kodak Digital Learning Center, www.kodak.com/US/en/digital/dlc/book3/chapter2/index.shtml, (acc. Oct 2000), (created 1994).

"COLLAGE-project by Guildhall Library." Corporation of London, collage.nhil.com/index.htm, (acc. Oct 2000), (created 1997).

"Color Management." Dunway Products, www.dunway.com/secties/Information/ColorManagement/ColorManagement.html, (acc. Jan 2000).

"Color Scanners: How We Tested." *PC Magazine Online*, no. March 4 (1997). www.zdnet.com/pcmag/features/scanners/hwt.htm, (acc. Oct 2000).

"Colour Characteristics of Digital Still Cameras (DSCs) Using Colour Targets and Spectral Illumination." *Program of Work,* PIMA/IT10: Electronic Still Picture Imaging, www.pima.net/standards/it10/IT10_POW.htm, (acc. Oct 2000), (created Sept 2000*).

"Conservation Implications of Digitization Projects." 1999, Library of Congress, memory.loc.gov/ammem/techdocs/conserv83199a.pdf, (acc. Oct 2000), (created Sept 1999).

"Creation of the Images and Database." Historic American Sheet Music, Rare Books, Manuscript and Special Collections Library, scriptorium.lib.duke.edu/sheetmusic/techinfo-scanning.html, (acc. Oct 2000), (created 1999).

"Data Attributes." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* lcweb.loc.gov:8081/ndlint/repository/attribs.html, (acc. Oct 2000), (created May 1997).

"Data Definitions." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* lcweb.loc.gov:8081/ndlint/repository/attdefs.html, (acc. Oct 2000), (created May 1997).

"Data Standards." Visual Resources Association, www.oberlin.edu/~art/vra/dsc.html, (acc. Oct 2000), (created June 2000).

"Digital Arts & Sciences/Gallery Systems home page." www.dascorp.com, (acc. Oct 2000), (created Nov 1999).

"Digital Camera Reviews." Digital Photography Reviews, www.inconference.com/digicam/camera.html, (acc. Oct 2000).

"Digital Image Basics: Some Explanations and Descriptions of Terms." *Digital Images in Multimedia Presentations,* Technical Advisory Service for Images (TASI), www.ilrt.bris.ac.uk/mru/dimp/section3/basics_prt.html, (acc. Oct 2000).

"Digital imaging and optical storage technology helps open up railway archives at National Railway Museum." *Electronic Library* 11, no. 2 (1993): 108-109.

"Digital Page Imaging and SGML: An Introduction to the Electronic Binding DTD (Ebind)." Berkeley Digital Library SunSITE, sunsite.berkeley.edu/Ebind/index.html, (acc. Oct 2000), (created June 1996).

"Electronic Imaging and Museums." *Visual Resources* 7 (1991).

"Electronic Library Image Service for Europe Project (ELISE 1 and 2)." nile.dmu.ac.uk/elise, (acc. Oct 2000), (created Feb 2000*).

"EmbARK, Image Management Module. Version 1.2. Review." *Visual Resources* XII (1996): 81-93.

"Fully Utilizing Photo CD Images: Typical Monitor Luminance for Photo YCC Luma Values - Article Number 6." Eastman Kodak Company, www.kodak.com/US/en/digital/techInfo/pcd-102.shtml, (acc. Jan 2000), (created Aug 1995).

Guidelines for Digital Imaging: Papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998. London: The National Preservation Office, 1998. www.rlg.ac.uk/preserv/joint/, (acc. Oct 2000).

"Higher Education Digitisation Service (HEDS)." heds.herts.ac.uk/, (acc. Oct 2000).

"Image Access homepage." www.imageaccess.com, (acc. Aug 2000), (created 1999).

"Image Capture and Manipulation: Guidelines and Procedures." Technical Advisory Service for Images (TASI), www.tasi.ac.uk/framework/capture/img-capt.html, (acc. Oct 2000), (created 2000*).

"Image Capture: Hardware and Software Summary." Technical Advisory Service for Images (TASI), www.tasi.ac.uk/framework/capture/hwandsw.html, (acc. Oct 2000), (created 2000).

"Image File Formats List." Center for Innovative Computer Applications, http://cica.indiana.edu/graphics/image.formats.html, (acc. Jan 2000), (created Jan 1994).

"Image Scanning." Harvard University, Library Digital Initiative, hul.harvard.edu/ldi/html/image_scanning.html,

(acc. Oct 2000), (created 1998).

“Images Collection: Introduction.” *Library and Information Centre,* The Royal Society of Chemistry, www.rsc.org/lic/imagesintro.htm, (acc. Oct 2000).

“Imaging & Document Solutions.” Imaging Magazine, www.imagingmagazine.com, (acc. Oct 2000).

“Indexing and Morphology: Why You Need Morphology.” Lingsoft Inc., www.lingsoft.fi/doc/indexing/morph.html, (acc. Oct 2000), (created Dec 1994).

“Information Architecture White Paper IA-6801: Electronic Image Formats and Compression Algorithms.” *Rev. 2.0,* 1998, Los Alamos National Laboratory, Information Architecture, www.lanl.gov/projects/ia/, (acc. Oct 2000).

“Instrumental Quality Control.” In *Fundamentals of Color and Appearance*, New Windsor, NY: MacBeth, 1996.

“Internet Library of Early Journals (January 1996 - August 1998). A Project in the eLib Programme: Final Report.” www.bodley.ox.ac.uk/ilej/papers/fr1999/fr1999.htm, (acc. Oct 2000), (created March 1999).

“KODAK PHOTO CD Technology/Transferring Images.” *Kodak Service & Support Frequently Asked Questions,* www.kodak.com/cluster/global/en/service/faqs/faq1001c.shtml, (acc. Oct 2000), (created Nov 1998*).

“Manchester Computerised Local Image Collection.” Manchester Local Studies Unit, www.manchester.gov.uk/mccdlt/lsuimage/home.htm, (acc. Oct 2000), (created Aug 2000*).

“Monitor Adjustment Target.” National Archives and Records Administration, NARA Archival Information Locator (NAIL), www.nara.gov/nara/target.html, (acc. Jan 2000), (created March 2000*).

“Munsell Color Vision Tests.” GretagMacbeth, www.munsell.com/munvis.htm, (acc. Jan 2000).

“NISO/CLIR/RLG Technical Metadata Elements for Images Workshop.” 1999, NISO (National Information Standards Organization), www.niso.org/image.html, (acc. Oct 2000), (created April 1999).

“OCR and Text Recognition: Commercial Research and Products.” Video and Image Data Exploration and Organization (VIDEO), hera.itc.it:3003/~messelod/OCR/Products.html, (acc. Oct 2000).

“Optical Image Technology, Inc. home page.” www.opticaltech.com, (acc. Oct 2000).

“PIMA/IT10: Electronic Still Picture Imaging.” International Organization for Standardization, http://mindit.netmind.com/searchit/track.shtml?URL=http://www.pima.net/standards/iso/tc42/WG18.htm&SOURCE=5, (acc. Oct 2000), (created July 2000*).

“Questions to Consider Before Beginning an Image Database Project.” University of Illinois Library, Digital Imaging & MediaTechnology Initiative, images.library.uiuc.edu/resources/20questns.html, (acc. Oct 2000), (created Dec 1999).

“Rare works digitised.” *Library Association Record* Technology Supplement, 96, no. 2, February (1994).

“Sample Scans: The Electronic Archive of Early American Fiction.” University of Virginia Library, etext.lib.virginia.edu/projects/scantest.html, (acc. Oct 2000).

“Scanners and Digital Cameras.” *RLG DigiNews* 1, no. 1, April 15 (1997). www.thames.rlg.org/preserv/diginews/diginews1.html, (acc. Oct 2000).

“Setting the Resolution.” *Photography Collection,* Denver Public Library, Western History/Genealogy Department, gowest.coalliance.org/calib.htm, (acc. Oct 2000).

“Specification ICC.11998-09, FileFormat for Color Profiles.” ICC Specifications, International Color Consortium, www.color.org/profiles.html, (acc. Oct 2000), (created Sept 1998).

“Technical Review: Outsourcing Film Scanning and Computer Output Microfilm (COM) Recording.” *RLG DigiNews* 1, no. 2, August (1997). www.rlg.org/preserv/diginews/diginews2.html#hardware&software, (acc. Oct 2000).

“The Aberdeen Bestiary: A pilot scheme for producing digitised images of manuscripts on the World Wide Web.” Aberdeen University, www.clues.abdn.ac.uk:8080/besttest/firstpag.html, (acc. Oct 2000).

"The American Museum of Photography." www.photographymuseum.com, (acc. Oct 2000), (created 2000).

"The CD-Information Center." www.cd-info.com, (acc. Oct 2000), (created 1999).

"The Conversion Process." Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html#conversion, (acc. Oct 2000), (created June 1999).

The effects of digitisation on library and information services: proceedings of a conference organised by NORDINFO and the British Library (Research and Development Department), held at the Holiday Inn, Edinburgh, Scotland, 17-20 September 1992, *British Library R & D report 6098*. 1993, London: British Library Research and Development Department.

The Secrets of Color Management. Randolph, MA: Agfa Educational Publishing, 1997.

"What is the Best OCR? OCR Software? Scanner Survey Results." hosc.net, www.hosc.net/ocr, (acc. Oct 2000), (created July 1997*).

"Workflow Guidelines: 2. Pre-scanning Preparation and Post-scanning Processing of the Originals." TASI, www.tasi.ac.uk/framework/collections/jidi-workflow.html#wg2, (acc. Oct 2000).

AGFA. *An Introduction to Digital Scanning.* Belgium: Agfa-Gevaert, 1994.

Allen, D. Y. "Creating and Distributing High Resolution Cartographic Images." *RLG DigiNews* 2, no. 4, August 15 (1998). www.rlg.org/preserv/diginews/diginews2-4.html#feature, (acc. Oct 2000).

American Society for Testing Materials. *ASTM Standards on Color and Appearance Measurement.* 5th ed. West Conshohocken, PA: ASTM, 1996.

ANSI/AIIM MS44-1988. *Recommended Practice for Quality Control of Image Scanners.* Silver Spring, MD: Association for Information and Image Management, 1988.

ANSI/AIIM TR26-1993, Resolution as it Relates to Photographic and Electronic Imaging. Silver Spring, MD: Association for Information and Image Management, 1993.

Ascher, R. N., and G. Nagy. "A means for achieving a high degree of compaction on scan-digitized printed text." *IEEE Transactions on Computers* C-23, no. 11, Nov (1974): 1174-1179.

Association for Information and Image Management. *Practice for Operational Procedures/Inspection and Quality Control of First-Generation, Silver Microfilm of Documents.* Silver Spring, MD: Association for Information and Image Management, 1998.

Association for Information and Image Management. *Selecting and Appropriate Image Compression Method to Match User Requirements.* Silver Spring, MD: Association for Information and Image Management International, 1998.

Barten, P. "MTF, CSF, and SQRI for Image Quality Analysis, short course notes." In *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology*, 1996.

Baxes, G. A. *Digital Image Processing.* New York: John Wiley & Sons, Inc., 1994.

Berger, B. "Digital Imaging Projects at Cornell University." In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 49-57.

Bescós Ramón, J. "Image Processing Algorithms for Readability Enhancement of Old Manuscripts." *Electronic Imaging* 89, no. 1 (1989): 392-97.

Bescós Ramón, J., F. Jaque, and L. Montoto. "Reflectance and Optical Contrast of Old Manuscripts: Wavelength Dependance." In *Scanning imaging: ECO1: 21-23 September 1988*, edited by Tony Wilson, 1989. 258-62.

Bescós Ramón, J., J. Navarro, and C. Ramón. "Mejora de Legibilidad de Documentos Antiguos Mediante Tratamiento Digital de Imágenes." In *IV Simposium Nacional de Reconocimiento de Formas y Análisis de Imágenes*, Granada: Sociedad Española de Reconocimiento de Formas y Análisis de Imágenes, 1990. 51-58.

Bescós Ramón, J., J. Pedro Secilla, and J. Navarro. "Filtering and Compression of Old Manuscripts by Adaptive

Processing Techniques." In *Proceedings of the Society for Information Display International Symposium 1990*, Las Vegas: Society for Information Display, 1990. 384-87.

Besser, H. "Best Practices for Image Capture." 1999, California Digital Library, sims.Berkeley.edu/~howard/moa2/bp90.doc, (acc. Oct 2000), (created July 1999).

Besser, H. "Digital Images for Museums." *Museum Studies Journal* 3, no. 1, Fall-Winter (1987): 74-81.

Besser, H., and J. Trant. *Introduction to Imaging: Issues in constructing an Image Database*. Getty Research Institute, 1996.

Bicknese, D. "Measuring the Accuracy of the OCR in the Making of America." *(a report prepared in fulfillment of directed field experience requirements, University of Michigan, School of Information, winter 1998),* University of Michigan, moa.umdl.umich.edu/moaocr.html, (acc. Oct 2000), (created 1998).

Birdsey, C. L. "Creating Accessible Digital Imagery." *D-Lib Magazine* 6, no. 4, April (2000). www.dlib.org/dlib/april00/birdsey/04birdsey.html, (acc. Oct 2000).

Black, K. "ELISE: an online image retrieval system." *ASLIB Information* 21, no. 7/8, July/August (1993): 293-295.

Blatner, D., and S. Roth. *Real World Scanning and Halftones.* Berkeley, CA: Peachpit Press, 1993.

Bone, J. "Color Scanning." *The Scanning FAQ,* Infomedia, www.infomedia.net/scan/TSF-Color.html, (acc. Oct 2000), (created July 1997).

Bone, J. "Greyscale Scanning." *The Scanning FAQ,* Infomedia, www.infomedia.net/scan/TSF-GreyScale.html, (acc. Oct 2000), (created July 1997).

Bone, J. "Line Art Scanning." *The Scanning FAQ,* Infomedia, www.infomedia.net/scan/TSF-LineArt.html, (acc. Oct 2000), (created July 1997).

Bone, J. "Scanning Resolution Calculator." *The Scanning FAQ,* www.infomedia.net/scan/calcs.html, (acc. Oct 2000), (created July 1997).

Booth, B., and D. Hopkin. "Catching the Trains: an application of high resolution digital imaging technology at the National Railway Museum, York." *Information Services and Use* 14, no. 3 (1994): 135-140.

Bouché, N. *Digitization for scholarly use: The Boswell Papers Project at the Beinecke Rare Book and Manuscript Library.* Washington, DC: Council on Library and Information Resources, 1999.

Braudaway, G. "Restoration of faded photographic transparencies by digital image processing." In *Proceedings of IS&T's 46th Annual Conference*, Springfield, VA: IS&T, 1993. 287-289.

Braunegg, D. J., R. D. Forkert, and N. B. Nill. *Rescaling Digital Fingerprints: Techniques and Image Quality Effects, MITRE Report for FBI Fingerprint Project.* 1995.

Brown, C. W., and B. J. Shephard. *Graphics File Formats, Reference and Guide.* Upper Saddle River, NJ: Prentice Hall PTR, 1995.

Brown, D. S. "Image Capture Beyond 24-Bit RGB." *RLG DigiNews* 3, no. 5, October 15 (1999). www.rlg.org/preserv/diginews/diginews3-5.html#technical, (acc. Oct 2000).

Bruemmer, B. H. "On the Bleeding Edge." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 122-124.

Burdick, H. E. *Digital Imaging: Theory and Applications.* New York: McGraw Hill, 1997.

Burns, P. D. "What Does the Colour Characterisation of Digital Still Cameras Standard Promise for the Digital Imaging Community?" In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 77.

Chambliss, M. A., J. A. Dawson, and E. J. Borg. "Measuring the MTF of Undersampled IRFA Sensors Using 2D Discrete Transformations." In *SPIE Proceedings: Infrared Imaging Systems, Design, Analysis, Modeling, and Testing VI*, edited by G C Holst, 1995. 312-324.

Chapman, S. “Guidelines for Image Capture.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 39-60.

Chapman, S., and A. R. Kenney. “Digital Conversion of Research Library Materials: A Case for Full International Capture.” *D-Lib Magazine* 2, no. 10, October (1996). www.dlib.org/dlib/october96/cornell/10chapman.html, (acc. Oct 2000).

Chapman, S., P. Conway, and A. R. Kenney. “Digital Imaging and Preservation Microfilm: The Future of the Hybrid Approach for the Preservation of Brittle Books.” CLIR, Commission on Preservation and Access, www.clir.org/cpa/archives/hybridintro.html, (acc. Oct 2000).

CIE Technical Committee 8-02. “Colour Difference Evaluation in Images.” International Commission on Illumination, www.colour.org/tc8-02, (acc. Oct 2000), (created Oct 1998).

Coles, T., A. Alexaner, and G. Shaw. “Following the Script: Optical Character Recognition Technology and the British Town and Trade Directory.” *History and Computing* 9 (1997): 1-16.

Colet, L. S., K. Keller, and E. Landsberg. “Digiting Photographic Collections: A Case Study at the Museum of Modern Art, NY.” *Spectra* 25, no. 2 (1997): 22-27.

Conway, P. *Preservation in the Digital World.* Washington, D.C.: Commission on Preservation and Access, 1996b.

Dahl, K. “OCR Trends and Implications.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 111-112.

Dale, R. “Lossy or Lossless? File Compression Strategies Discussion at ALA.” *RLG DigiNews* 3, no. 2, February 15 (1999). www.rlg.org/preserv/diginews/diginews3-1.html#technical1, (acc. Oct 2000).

D’Amato, D. P., and R. C. Klopfenstein. “Requirements and Options for the Digitization of the Illustration Collections of the Natural Museum of Natural History.” 1996, www.nmnh.si.edu/cris/techrpts/imagopts, (acc. Oct 2000), (created March 1996).

Davis, B. “The Digital Museum and Computer-Generated Collections and Catalogs of Art.” *Aperture* 136, no. Summer (1994): 68-70.

Department of Preservation. “Cornell Online Imaging Tutorial.” Cornell University, www.library.cornell.edu/preservation, (acc. Oct 2000), (created Feb 1998*).

Donovan, K. “Anatomy of an Imaging Project: A Primer for Museums, Libraries, Archives and Other Visual Collections.” *Spectra, A Publication of the Museum Computer Network* 23, no. 2, Winter (1995-6): 19-22.

Donovan, K. “The Promise of FlashPix™ Image File Format.” *RLG DigiNews* 2.2, no. April 15 (1998). www.rlg.org/preserv/diginews/diginews22.html#FlashPix, (acc. Oct 2000).

Downton, A. C., and S. Impedovo, (eds.) *Progress in handwriting recognition*. 1997, Singapore: World Scientific.

Elkington, N., (ed.) *RLG Preservation Microfilming Handbook*. 1992, Mountain View, CA: Research Libraries Group.

Engel, V. S. “The Browser Safe Palette (BSP).” www.the-light.com/netcol.html, (acc. Oct 2000), (created 1996).

Ester, M. “Building Towards Image Resources and User Environments in the Visual Arts.” In *Proceedings of the 1993 Electronic Imaging and the Visual Arts Conference: EVA ’93*, edited by James Hemsley, Neil Sandford and David Saunders, Aldershots, Hants: Braumer Ltd., 1993.

Ester, M. “Digital Images in the Context of Visual Collections and Scholarship.” *Visual Resources* 10, no. 1 (1994): 11-24.

Ester, M. “Image Quality and Viewer Perception.” *Leonardo* 23, no. 1 (1990): 51-63.

Ester, M. “Image Use in Art - Historical Practice.” In *Gateways, Gatekeeers, and Roles in the Information Omniverse, Proceedings form the Third Symposium*, edited by Ann Okerson and Dru Mogge, Washingon, DC:

Association of Research Libraries - Association of American University Presses, 1993.

Ester, M. “Issues in the Use of Electronic Images for Scholarship in the Arts and Humanities.” In *Networking in the Humanities (proceedings of the Second Conference on Scholarship and Technology in the Humanities, held at Elvetham Hall, Hampshire, UK, 13-16 April, 1994: papers in honour of Michael Smethurst for his 60th birthday)*, edited by S. Kenna and S. Ross, London: Bowker-Saur, 1995. 111-125.

Ester, M. “Specifics of Imaging Practice.” *Archives & Museum Informatics* 9 (1995): 147-185.

Evers, D. E., G. O. Holm, and C. E. Sasser. *Digitization of the floating marsh maps in the Barataria and Terrebonne basins, Louisiana.* Thibodaux, LA: Barataria-Terrebonne National Estuary Program, 1996.

Eyre, J. “ELISE II: Networked Delivery of Images and User Interface Issues.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 31-34.

Federal Task Force on Digitization (Canada). *Towards a learning nation: The digital contribution: recommendations proposed by the Federal Task Force on Digitization: final report.* Ottawa: National Library of Canada, 1999.

Fischer, T. A., and J. Holm. “Electronic Still Picture Camera Spatial Frequency Response Measurement.” In *Proceedings, IS&T 47th Aunnual Conference*, Rochester, NY, 1994. 626-630.

Flickner, M., H. Swahney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. “Query by Image and Video Content: The QBIC System.” *IEEE Computer* 28, no. 9 (1995): 23-32.

Forkert, R. D., and et al. *Test Procedures for Verifying IAFIS Scanner Image Quality Requirements, MITRE Report for FBI Fingerprint Project.* 1994.

Fraser, B. “Out of Gamut: Stop the sRGB Bandwagon.” *MacWeek*, no. July 31 (1998). macweek.zdnet.com/1229/op_fraser.html, (acc. Oct 2000).

Frey, F. “Benchmarking Digitizing Systems for Photographs.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 36.

Frey, F. “Digital Imaging for Photographic Collections: Foundations for Technical Standards.” *RLG DigiNews* 1, no. 3, December 15 (1997). www.rlg.org/preserv/diginews/diginews3.html#com, (acc. Oct 2000).

Frey, F. “Digitization of Photographic Collections.” In *Proceedings, IS&T 50th Annual Conference*, Boston, 1997. 597-599.

Frey, F. S., and J. M. Reilly. *Digital Imaging for Photographic Collections: Foundations for Technical Standards.* Rochester, NY: Image Permanence Institute, Rochester Institute of Technology, 1999. www.rit.edu/~661www1/sub_pages/page17.pdf, (acc. Oct 2000).

Frey, F., and S. Süsstrunk. “Color Issues to Consider in Pictorial Image Databases.” In *Proceedings, IS&T Fifth Color Imaging Conference*, Scottsdale, AZ, 1997. 112-115.

Frey, F., and S. Süsstrunk. “Image Quality Issues for the Digitization of Photographic Collections.” In *Proceedings, IS&T 49th Annual Conference*, Minneapolis, MN, 1996. 349-353.

Frey, F., R. Gschwind, and L. Rosenthaler. “Electronic Imaging, a Tool for the Reconstruction of Faded Color Photographs and Motion Pictures.” In *Proceedings, IS&T Fourth Color Imaging Conference*, 1996. 39-44.

Fulton, W. “A Few Scanning Tips.” www.scantips.com, (acc. Oct 2000), (created 1997).

Gagos, T., with, P. Daub, A. Loftus, and S. Zachary. “Scanning the Past: A Modern Approach to Ancient Culture.” *Library Hi Tech* 14, no. 1 (1996): 11-22.

Gann, R. G. *Desktop Scanners: Image Quality Evaluation.* Upper Saddle River, NJ: Prentice-Hall, 1999.

Gann, R., J. Goode, and R. Wadas. *Reviewing and Testing Desktop Scanners.* Second ed. Greeley, CO: Hewlett Packard Company, 1995.

Gertz, J. “Guidelines for Digital Imaging.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 11-20.

Gertz, J. *Oversize Color Images Project, 1994-1995, A Report to the Commission on Preservation and Access.* Washington, DC: Commission on Preservation and Access, 1995. www.columbia.edu/dlc/nysmb/reports/phase1.html,

Getz, M. “Questionnaire for Digital Scriptorium Users.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSquery.pdf, (acc. Oct 2000), (created March 1998).

Getz, M. “The Digital Scriptorium: Medieval Manuscripts on the World Wide Web.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSvision.pdf, (acc. Oct 2000), (created March 1998).

Gill, T., C. Grout, and L. Smith. “Visual arts, museums and cultural heritage information standards: a domain-specific review of relevant standards for networked information discovery.” Visual Arts Data Service, vads.ahds.ac.uk/training_advice/standards/standards_review/info%20standards.html, (acc. Oct 2000), (created March 1997).

Gillings, M., and A. Wise, (eds.) *GIS Guide to Good Practice*. 1999, Oxford: Oxbow Books for the Arts and Humanities Data Service.

Giorgianni, E. J., and T. E. Madden. *Digital Color Management.* Reading, MA: Addison-Wesley, 1997.

Gordon, C. “An Introduction to ICONCLASS.” In *Terminology for Museums: Proceedings of an International Conference Held in Cambridge, England, 21-24 September 1988*, edited by D Andrew Roberts, Cambridge, UK: Museum Documentation Association with the Assistance of the Getty Grant Program, 1990. 233-244.

Gordon, M. E. “Image Databases in North America.” *Special issue of Computers in the History of Art* 4, no. 2 (1994).

Greenberg, A. D. *Digital images: a practical guide.* Berkeley, CA: Osborne McGraw-Hill, 1995.

Helmuth, N. “Digital Photography.” www.digital-photography.org/default.html, (acc. Oct 2000), (created Dec 1999*).

Holm, J. “Factors to Consider in Pictorial Digital Image Processing.” In *Proceedings, IS&T 49th Annual Conference*, Minneapolis, MN, 1996. 298-304.

Holm, J. “Log NEQ Based Pictorial Print Noise Characterization.” In *Proceedings, IS&T 47th Annual Conference*, Rochester, NY, 1994. 634-640.

Holm, J. “Survey of Developing Electronic Photography Standards.” *Critical Reviews of Optical Science and Technology, SPIE* CR61 (1996): 120-152.

Holm, J. “The Evaluation of Digital Photography Systems, Short Course Notes.” In *IS&T 49th Annual Conference*, 1996.

Holm, J., and N. Judge. “Electronic Photography at the NASA Langley Research Center.” In *Proceedings, IS&T 48th Annual Conference*, Washington, DC, 1995. 436-441.

Hopkin, D. “Shifting the Focus: Digital Imaging & Photographic Collections Management at the National Railway Museum.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 72-80.

Howard, P. G., F. Kossentini, B. Martins, S. Forchhammer, and W. Rucklidge. “The Emerging JBIG2 Standard.” *IEEE Transactions on Circuits and Systems for Video Technology* 8, no. 7, Nov (1998): 838-848.

Howell, A. “Film scanning from newspaper collections: International initiatives.” *RLG DigiNews* 1, no. 2 (1997). www.rlg.org/preserv/diginews/diginews2.html#film-scanning, (acc. Oct 2000).

Hüfner, J. “Steigerung der Erkennungsgenauigkeit durch maschinellen Abgleich verschiedener durch OCR erzeugter Volltexte (Increasing the Quality of OCR-generated Full Texts by Automatic Comparison).” *Nachrichten für Dokumentation* 48, no. 2 (1997): 79-85.

International Organization for Standardization. *Graphic Technology - Prepress Digital Data Exchange - Tag Image File Format for Image Technology (TIFF/IT).*, *Technical Committee 130, ISO/FDIS 12639.* Geneva: Internation Organization for Standardization, 1998.

ISO 12233/FDIS, ISO/TC42, Photography - Electronic Still Picture Cameras - Resolution Measurements. 1999.

ISO 14524/FDIS, ISO/TC42, Photography - Electronic Still Picture Cameras - Methods for Measuring Opto-Electronic Conversion Functions (OECFs). 1999.

ISO 15739/CD, ISO/TC42, Photography - Electronic Still Picture Cameras - Noise Measurements. 1999.

ISO 16067/WD, ISO/TC42, Electronic scanners for photographic images - Spatial resolution measurements: Part 1, Scanners for reflective media. 1999.

ISO 17321/WD, ISOTC42, Graphic Technology and Photography - Colour Characterisation of Electronic Still Cameras (DSCs) Using Colour Targets and Spectral Illumination. 1999.

ISO/DIS 12234-2, Photography - Electronic Still Pictures Cameras - Removable Memory - Part 2: Image Data Format - TIFF/EP. 1998.

ISO/FDIS 12233 Photography, Electronic Still Picture Cameras, Resolution Measurements.

ISO/IEC. “IS 12087-5, Information Technology - Computer Graphics and Image Processing - Image Processing Interchange (IPI) - Functional Specification - Part 5: Basic Image Interchange Format (BIIF).” 1998, http://164.214.2.51/ntb/related/, (acc. Jan 2000), (created Dec 1998).

Jenkins, F. R., T. A. Nartker, and S. V. Rice. “Testing OCR Accuracy: Results of the Fifth Annual Test of OCR Technology by UNLV’s Information Science Research Institute.” *INFORM* September 10, no. 8 (1996): 20-25.

Kenney, A. R. “Conversion of Traditional Source Materials into Digital Form.” In *Research Agenda for Networked Cultural Heritage*, edited by David Bearman, Santa Monica, CA: Getty Art History Information Program, 1996. 41-47.

Kenney, A. R. “Digital Benchmarking for Conversion and Access.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 24-60.

Kenney, A. R., and L. H. Sharpe, II. “Illustrated Book Study: Digital Conversion Requirements of Printed Illustrations.” 1999, Library of Congress, Preservation, www.lcweb.loc.gov/preserv/rt/illbk/ibs.htm, (acc. Oct 2000), (created July 1999).

Kenney, A. R., and O. R. Rieger. “Using Kodak Photo CD Technology for Preservation and Access.” 1998, Cornell University Library Preservation & Conservation, Ithaca, NY, www.library.cornell.edu/preservation/kodak/cover.htm, (acc. Oct 2000), (created May 1998).

Kenney, A. R., and S. Chapman, (eds.) *Digital Imaging for Libraries and Archives*. 1996, Ithaca, NY: Cornell University Library Press.

Kenney, A. R., and S. Chapman. *Tutorial. Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality.* Washington, DC: Commission on Preservation and Access, 1995.

Kenney, A., and S. Chapman. *Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality.* Washington, DC: Commission on Preservation and Access, 1995.

Kessler, B. R. “Electronic Images in Visual Resources Collection: Some Strategic Question.” *Visual Resources* X (1994): 1-10.

Kinoshita, H., and T. Yamamuro. “Effects of Repetitive JPEG Compressions with DCT Block Rearrangement Operation on Image Quality.” *Journal of Imaging Science and Technology* 39, no. 6, November/December (1995): 546-558.

Kirtley, T. “Preservation and digitisation: principles, practice and policies: The 1996 National Preservation Office Conference.” *Library Conservation News* 53, no. Winter (1997).

Klak, J. “Pretty pixels: digital imaging in Kirkcaldy.” *The Local Studies Librarian* 16, no. 2(3) (1997): 1-9.

Klijn, E., and Y. de Lusenet. *In the Picture: Preservation and Digitisation of European Photographic Collections.* Amsterdam: European Commission on Preservation and Access, 2000.

Knoll, A. “Digitisation of Rare Library Materials: Compound Documents - Manuscripts and Old Printed Books.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 23-25.

Knoll, A., and S. Psohlavec. *Digitization of old manuscripts.* Praha: Narodni knihovna, 1995.

Lawler, B. *Optimizing Photo CD scans for Prepress and Publishing.*: Eastman Kodak company, 1995. ftp://ftp.kodak.com/pub/photo-cd/prepress/dci346.pdf, (acc. Oct 2000).

Leith, L., and J. Monk. *Digitization of the book: a report on present trends.* Toronto: Canadian National Institute for the Blind, Library for the Blind, 1996.

Limp, F. W. *The Use of Multispectral Digital Imagery in Archaeological Investigations.*, *Report to US Army Corp of Engineers, Arkansas Survey Research Series 34.* Fayetteville, Arkansas 1989.

Lossau, N. “Digitisation - technical issues: Production at the Göttingen Digitisation Center (GDZ).” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 47-51.

Lossau, N., and F. Klaproth. “Digitization Efforts at the Center for Retrospective Digitization, Göttingen University Library.” *RLG DigiNews* 3, no. 1, February 15 (1999). www.rlg.org/preserv/diginews/diginews3-1.html#feature2, (acc. Oct 2000).

Lossau, N., and F. Klaproth. “TIFF Header: A Reference Stamp for Image Files.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 97-98.

Macklin, L. L., and S. L. Lockmiller. *Digital Imaging of Photographs. A Practical Approach to Workflow Design and Project Management.*, *LITA Guides no. 4.* Chicago: American Library Association, 1999.

Maitre, H., F. J. M. Schmitt, and J. Crettez. “High Quality Imaging in Museums from Theory to Practice.” In *Very High Resolution and Quality Imaging II, Proceedings SPIE*, edited by V Algazi, S Ono and A G Tescher: SPIE, 1997. 30-39.

Mantelman, L. “Windows OCR Packages Put a Typist in Your Tank.” *Imaging Magazine* December (1994): 8-25.

Martinez, K. “High Resolution Digital Imaging of Paintings: The VASARI Project.” *Microcomputers for Information Management* 8, no. 4, December (1991): 277-83.

Mayer, M. “Scanners.” *Message posted to Conservation DistList, 9 December 1997,* 1997, palimpsest.stanford.edu/byform/mailing-lists/cdl/1997/1549.html, (acc. Oct 2000), (created Sept 1997).

McCamy, C. S., H. Marcus, and J. G. Davidson. “A Color Rendering Chart.” *Journal of Applied Photographic Engineering* 2 (1976): 95-99.

McElhone, J. “Electronic Imaging: Information Gathered at Two Events.” *Spectra* 25, no. 1 (1997): 48-56.

McTavish, S. “DoD-NARA Scanned Images Standards Conference.” *RLG DigiNews* 3, no. 2, April 15 (1999). www.rlg.org/preserv/diginews/diginews3-2.html#technical1, (acc. Oct 2000).

Means, L. “Digitisation as Transformation and Some Implications for the Arts.” *Leonardo* 17, no. 3 (1984): 195-199.

Mecocci, A. “Practical Experiences and Techniques Used for Acquiring Images of Large Historical Buildings and Cultural Items.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 13-14.

Mehrabi, H., and H. Laursen. “Standards for images and full text.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 43-46.

Mitchell, W. J. *The Reconfigured Eye: Visual Truth in the Past-Photographic Era.* Boston: Massachusetts

Institute of Technology, 1992.

Myaeng, S., D. Shin, W. Chun, and K. Lee. “MIRAGE: A prototype for a multimedia digital library.” *Journal of the Korean Information Science Society* 3, no. 6, December (1997): 605-618.

Nartker, T. A., and S. V. Rice. “OCR Accuracy: UNLV’s Third Annual Test.” *INFORM* 8, no. 8 (1994): 30-36.

Nartker, T. A., S. V. Rice, and J. Kanai. “OCR Accuracy: UNLV’s Second Annual Test.” *INFORM* 8, no. 1 (1994): 40-45.

Netherlands Historical Data Archive, and Nijmegen Institute for Cognition and Information, (eds.) *Optical character recognition in the historical discipline: proceedings of an international workshop*. Leiden 10-12 June 1993 ed. Vol. 18, *Halbgraue Reihe zur historischen Fachinformatik. Serie A, Historische Quellenkunden*. 1993, Göttingen: Max-Planck-Institute für Geschichte.

Nill, N. B., and B. R. Paine. *A Computer Program to Determine the Sine Wave Modulation Transfer Function of Image Scanners, MITRE Report for FBI Fingerprint Project.* 1994.

North Carolina, General Assembly, Legislative Research Commission, and Committee on Digitization of Public Records by State Archives. *Digitization of public records by state archives: report to the 2000 session of the 1999 General Assembly of North Carolina.* Raleigh, NC: The Commission, 2000.

Okon, C. “IBMs Image Recognition Technology for Databases at Work: QBIC or Not QBIC?” *Advanced Imaging* May (10) (1995): 63-65.

Ostrow, S. E. *Digitizing Historical Pictorial Collections for the Internet.* Washington: DC: CLIR, 1998. www.clir.org/cpa/reports/ostrow/pub71.html,

Parry, D. *Virtually New. Creating the Digital Collection. A Review of Digitisation Projects in Local Authority Libraries and Archives.* London: Library and Information Commission, 1998.

Podmore, H. “The Digitisation of Microfilm.” In *Towards the Digital Library: The British Library’s ‘Initiatives for Access’ Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott, London: The British Library, 1998. 68-72.

Poe, R. “Aesthetic Considerations in Tone and Color Management.” In *Proceedings, IS&T Third Color Imaging Conference*, Scotsdale, AZ, 1995. 164-168.

Poynton, C. “Frequently-Asked Questions About Gamma.” www.inforamp.net/~poynton/GammaFAQ.html, (acc. Oct 2000), (created Aug 1998).

Poynton, C. “Poynton’s Color FAQs.” www.inforamp.net/~poynton/ColorFAQ.html, (acc. Oct 2000), (created April 2000).

Prescott, A., and M. Pratt. “Excalibur: Image-Based Text Storage and Searching.” In *Towards the Digital Library: The British Library’s ‘Initiatives for Access’ Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott, London: The British Library, 1998. 178-189.

Price-Wilkin, J. “Access to Digital Image Collections: System Building and Image Processing.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 101-118.

Psohlavec, S. “Digitization of Manuscripts in the National Library of the Czech Republic.” *Microform and Imaging Review* 27, no. 1, Winter (1998): 21.

Puglia, S. “Fractal and Wavelet Compression.” *RLG DigiNews* 2, no. 3 (1998). www.rlg.org/preserv/diginews/diginews23.html#technical2, (acc. Oct 2000).

Rabbani, M. “Image Compression Fundamentals.” In *The Compression Experience: Proceedings of The Rochester Section of the SMPTE (Society of Motion Picture & Television Engineers) Tutorial*, Rochester, NY, 1995. 7-24.

Reichenbach, S. E., S. K. Park, and R. Naraynswamy. “Characterizing Digital Image Acquisition Devices.” *Optical Engineering* 30, no. 2 (1991): 170-177.

Reid, N. H. “Photographic Archives: Aberdeen, Dundee and St Andrews.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 106-119.

Reilly, B. F., J. M. Reilly, D. H. Norris, S. T. Puglia, P. De Stefano, J. Van Haaften, and P. McClung. *Photograph Preservation and the Research Library.* Edited by Jennifer Porro. Mountain View, CA: The Research Libraries Group, 1991.

Resolution as it Relates to Photographic and Electronic Imaging. ANSI/AIIM TR26-1993, *Technical Report.* Silver Spring, MD: Association for Information and Image Management, 1993.

Rhyne, C. S. “Computer Image for Research, Teaching, and Publication in Art History and Related Disciplines.” *Visual Resources* XII (1996): 19-51.

Ritter, N. “The Unofficial TIFF Homepage.” home.earthlink.net/~ritter/tiff/, (acc. Oct 2000), (created Feb 1997*).

Roberts, H. E. ""Second Hand Images: The Roles of Surrogates in Artistic and Cultural Exchange". Paper delivered as the Keynote Address for the Visual Resources Satellite Session of the XXVIIIth International Congress of the History of Art (CIHA), Berlin, July 1992.” *Visual Resources* 9, no. 4 (1993): 335-346.

Robinson, P. “Image Capture and Analysis.” In *New Technologies for the Humanities*, edited by Christine Mullings, Marilyn Deegan, Seamus Ross and Stephanie Kenna, London: Bowker Saur, 1996. 47-64.

Robinson, P. *The Digitization of Primary Textual Sources.* Oxford: Office for Humanities Communication Publications, No. 4, 1993.

Ross, S. “From conventional photographs to Digital Resources.” *Archealogical Computing Newsletter* 35 (1993): 14-21.

Russ, J. C. *The Image Processing Handbook.* second ed. Boca Raton, FL: CRC Press, 1995.

Schantz, H. F. “Optical Character Recognition - The Mature Technology with the Brilliant Future.” *INFORM* July/August 10, no. 7 (1996).

Seales, W. B., J. Griffioen, and K. Kiernan. “The Digital Atheneum - Restoring Damaged Manuscripts.” *RLG DigiNews* 3, no. 6, December 15 (1999). www.rlg.org/preserv/diginews/diginews3-6.html#technical1, (acc. Oct 2000).

Seaman, D. “Image Scanning: A Basic Helpsheet.” University of Virginia, Electronic Text Center, etext.lib.virginia.edu/helpsheets/scanimage.html, (acc. Oct 2000).

Sharpe, L. “Preservation-Quality Scanning of Bound Volumes: Integration of the Picture Elements ISE Board with the Minolta PS-3000 Book Scanner.” *RLG DigiNews* 1.1, no. April 15 (1997). http://www.rlg.org/preserv/diginews/diginews1.html#feature, (acc. Oct 2000).

Skaggs, D. “Electronic Document Management Systems as Archival Image Repositories.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 131-132.

Steele, M., J. Murdoch, and M. Arnott. “The Emergence of Photo-CD as a Preferred Method for Image Capture and Storage.” *Information Services and Use* 16 (1996): 9.

Stokes, J. R. “Imaging Pictorial Collections at the Library of Congress.” *RLG DigiNews* 3, no. 2, April 15 (1999). www.rlg.org/preserv/diginews/diginews3-2.html#feature, (acc. Oct 2000).

Stokes, J. R. “Sheet Music Scanning Procedures.” Historic American Sheet Music, Duke University, Rare Book, Manuscript and Special Collections Library, http://scriptorium.lib.duke.edu/sheetmusic/procedure.html, (acc. Oct 2000).

Sunderland, J. “Electronic Imaging and the Visual Arts: a Review of the Conference Held at Imperial College, London, July 23-27 1990.” *Computers and the History of Art* 1, no. 1 (1991): 55-61.

Sundt, C. L., (ed.),. “Issues in Electronic Imaging.” *Special issue of Visual Resources* 10, no. 1 (1994).

Swartzell, A. G. “Preparation of Materials for Digitisation.” In *Guidelines for Digital Imaging: papers given at the*

Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998, London: The National Preservation Office, 1998. 29-32.

Tanner, S. “Appendix J: Costed Collections.” In *Scoping the Future of The University of Oxford’s Digital Library Collections*, 1999.

Tanner, S., and J. L. Smith. “Digitisation: How Much Does it Really Cost?” *(paper for the Digital Resources for the Humanities 1999 Conference, Kings College London, September 12-15, 1999),* 1999, heds.herts.ac.uk/resources/papers/drh99.pdf, (acc. Oct 2000), (created Sept 1999).

TASI (Technical Advisory Service for Images). “TASI Photographic Guidelines.” TASI, www.tasi.ac.uk/building/photoguide1.html, (acc. Oct 2000).

UNESCO. “Digitizing Library and Archives Collections.” WebWorld, unesco.org/webworld/highlights/digitisation_0199.html, (acc. Oct 2000).

University of Virginia Library. “Producing Digital Images.” *The Electronic Archive of Early American Fiction: Comparing Usage and Costs Between Electronic Texts and Original Print Editions of Rare Early American Fiction,* www.lib.virginia.edu/speccol/mellon/image.html, (acc. Oct 2000), (created Feb 1999).

van Horik, R. “Recent progress in the automatic reading of printed historical documents.” In *History and Computing 5*, edited by P K Doornvan and R van Horik, 1993. 68-73.

Vitale, T. “Light Levels Used in Modern Flatbed Scanners.” *RLG DigiNews* 2, no. 5, 15 October (1998). www.rlg.org/preserv/diginews/diginews2-5.html#technical, (acc. Oct 2000).

Wallace, J. “Project Chapman: The Direct Delivery of Digital Smithsonian Photographic Images Via the Internet.” *Visual Resources* 10, no. 1 (1994): 57-60.

Wasner, R. *Display Subsystems for Production Imaging.* Hingham, MA: Rheinner, 1998.

Wayner, P. “Optical Character Recognition: an inside look at how optical character recognition works.” *Byte* December 18, no. 13 (1993): 203-120.

Weibel, S., and E. Miller. “Image Description on the Internet.” *D-Lib Magazine* 3, no. 1, January (1997). www.dlib.org/dlib/january97/oclc/01weibel.html, (acc. Oct 2000).

Weinstein, W. “Designing an Image Database: A Holistic Approach.” *Visual Resources* 10, no. 1 (1994): 49-56.

Weise, J. “DMS Guide to Web Color: Palettes and Gamma.” OIT Digital Media Solutions, www.oit.itd.umich.edu/projects/DMS/answers/colorguide, (acc. Oct 2000).

Weise, J. “Image Database Access Architecture.” *(version I.I., draft I),* 1997, University of Michigan, images.umdl.umich.edu/info/arch/Image_Tech_Dev_Plan.html, (acc. Oct 2000), (created June 1997).

Weise, J. “Overview of Media Database Issues and Technologies.” *OIT Digital Media Solutions,* www.oit.itd.umich.edu/projects/DMI/answers/pickmediadb.html, (acc. Jan 2000).

Weise, J. P. “Data Transfer Time.” *Preparing Quality Images for Computer Networks,* University of Michigan, www-personal.umich.edu/~jweise/quality/DataTransferTime.HTML, (acc. Oct 2000).

Wenman, L. “The Browser-Safe Color Palette.” www.lynda.com/hex.html, (acc. Oct 2000).

Williams, D. “An Overview of Image-Quality Metrics.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 72-73.

Williams, D. “Benchmarking of rhe ISO 12233 Slanted-edge Spatial Frequency Response Plug-in.” In *Proceedings, IS&T 1998 PICS Conference*, Portland, OR, 1998. 133-137.

Williams, D. “Image Quality Metrics.” *RLG DigiNews* 4, no. 4, August (2000). www.rlg.org/preserv/diginews/diginews4-4.html#technical1, (acc. Oct 2000).

Williams, D. “Interpreting Digital Scanner/Camera Specifications.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 40-41.

Williams, D. “What is an MTF...and Why Should You Care?” *RLG DigiNews* 2, no. 1, February 15 (1998). www.rlg.org/preserv/diginews/diginews21.html#technical, (acc. Oct 2000).

Williams, J. “Guidelines for Image Capture: a UK Perspective.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 61-71.

Wong, H. S. “Effect of Knife-Edge Skew on Modulation Transfer Function Measurements of Charge-Coupled Device Imagers Employing a Scanning Knife Edge.” *Optical Engineering* 30, no. 9 (1991): 1394-1398.

Woodward, H. “Selecting journals for digitisation: piecing together the puzzle to create a European model.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 77-83.

Wu, Y. “Raster, Vector, and Automated Raster-to-Vector Conversion.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 117-118.

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

5. Audio (capture; management)

"Digital Sound." Virginia Polytechnic Institute and State University, ei.cs.vt.edu/~netinfo/notes/chap1/digitalsound/digsnd.html, (acc. Oct 2000), (created 1998).

"Music for the Nation - Digitizing the Collection." Historic American Sheet Music, Rare Books, Manuscript and Special Collections Library, www.memory.loc.gov/ammem/smhtml/smtech.html, (acc. Oct 2000), (created Sept 1999).

Bainbridge, D., and T. C. Bell. "An extensible optical music recognition system." In *Proceedings of the 1996 Australasian Computer Science Conference*, Melbourne, Australia, 1996. 308-317.

Harris, M.N. and R. Kelly, D.A.McLeod, M.J.Story. "Effects in High Sample Rate Audio Material." DCS Ltd, 1998.

Johan, Lilkencrants. "Speech Signal Processing*." The Handbook of Phonetic Sciences*. Eds. Hardcastle William and Laver John. Oxford: Blackwell, 1997.

Karl, J.H. *An Introduction to Digital Signal Processing*. Academic Press, 1989.

Myaeng, S., D. Shin, W. Chun, and K. Lee. "MIRAGE: A prototype for a multimedia digital library." *Journal of the Korean Information Science Society* 3, no. 6, December (1997): 605-618.

Pohlmann, Ken. *Principles of Digital Audio*. New York: McGraw-Hill, 2000.

Rabiner, Lawrence R., and Ronald W. Schafer. *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice-Hall, 1978.

SONY Corporation. "Minidisk Specifications". Digital Audio Disk Corporation, 1993.

Stevens, Kenneth N. *Acoustic Phonetics*. Cambridge, Mass.: MIT Press, 1998.

Story, Mike. "A Suggested Explanation for (Some of) the Audible Differences between High Sample Rate and Conventional Sample Rate Audio Material". DCS Ltd, 1997.

Story, Mike and R. Kelly, D.A. McLeod. "Resolution, Bits, SNR and Linearity". DCS Ltd, 1998.

Titze, Ingo. *Workshop of Acoustic Voice Analysis*. National Center for Voice and Speech, 1994.

Vanderkooy, J., and S.P. Lipshitz. "Dither in Digital Audio". JAES, vol.32, no. 11 (1984).

Weise, J. "Overview of Media Database Issues and Technologies." *OIT Digital Media Solutions,* www.oit.itd.umich.edu/projects/DMI/answers/pickmediadb.html, (acc. Jan 2000).

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

6. Moving Images (capture; management)

"Association of Moving Image Archivists", www.amianet.org, (acc. Oct 2000).

"Digital Television Consumer Information." 1998, FCC, Office of Engineering and Technology, www.fcc.gov/Bureaus/Engineering_Technology/Factsheets/dtv9811.html, (acc. Oct 2000), (created Nov 1998).

"PRESTO: Preservation Technology for European Broadcast Archives", http://presto.joanneum.ac.at/index.asp, (acc. March 2002).

Chang, S.-F., J. R. Smith, H. J. Meng, H. Wang, and D. Zhong. "Finding Images/Video in Large Archives: Columbia's Content-Based Visual Query Project." *D-Lib Magazine* 3, no. 2, February (1997). www.dlib.org/dlib/february97/columbia/02chang.html, (acc. Oct 2000).

Chang, S.-F., W. Chen, H. J. Meng, H. Sundarm, and D. Zhong. "VideoQ: and automated content based video search system using visual cues." In *Proceedings of the ACM Conference on Multimedia '97*, 1997. 313-324.

Christel, M. G. "Informedia Goes to School: Early Findings from the Digital Video Library Project." *D-Lib Magazine* 2, no. 9, September (1996). www.dlib.org/dlib/september96/informedia/09christel.html, (acc. Oct 2000).

Fenske, D. E., and J. W. Dunn. "The VARIATIONS Project at Indiana University's Music Library." *D-Lib Magazine* 2, no. 6, June (1996). www.dlib.org/dlib/june96/variations/06fenske.html, (acc. Oct 2000).

Flickner, M., H. Swahney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. "Query by Image and Video Content: The QBIC System." *IEEE Computer* 28, no. 9 (1995): 23-32.

Frey, F., R. Gschwind, and L. Rosenthaler. "Electronic Imaging, a Tool for the Reconstruction of Faded Color Photographs and Motion Pictures." In *Proceedings, IS&T Fourth Color Imaging Conference*, 1996. 39-44.

Gilmour, I. "Digital audio in archives." *IASA Journal* 4, no. November (1994): 38-51.

Green, D. "Beyond Word and Image. Networking Moving Images: More Than Just the "Movies"." *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/07green.html, (acc. Oct 2000).

Gschwind, R., L. Rosenthaler, and F. Frey. "The Use of Digital Imaging for the Preservation of Collections of Photographs and Motion Pictures." In *Proceedings, ICOM 11th Triennial Meeting, Edinburgh, September 1-6*, 1996. 8-15.

Hauptmann, A. G., M. J. Witbrock, and M. G. Christel. "News-on-Demand: An Application of Informedia® Technology." *D-Lib Magazine* 1, no. 9, September (1995). www.dlib.org/dlib/september95/nod/09hauptmann1.html, (acc. Oct 2000).

Hunter, J. "A Proposal for an MPEG-7 DDL." In *P547, MPEG-7 AHG Test and Evaluation Meeting, Lancaster, February 1999*, 1999.

Hunter, J. “MPEG-7: Behind the Scenes.” *D-Lib Magazine* 5, no. 9, September (1999). www.dlib.org/dlib/september99/hunter/09hunter.html, (acc. Oct 2000).

MPEG-7 DDL Group. *MPEG-7 Description Definition Language Document V 1.0.*, *July 1999 ISO/IEC SC29/WG11/ N2862.* Vancouver 1999.

MPEG-7 Description Scheme Group. *MPEG-7 Description Schemes (V0.5).*, *July 1999 ISO/IEC SC29/WG11 N2844.* Vancouver 1999.

MPEG-7 Requirements Group. *MPEG-7 Requirements Document V.9.*, *July 1999 ISO/IEC SC29/WG11 N2859.* Vancouver 1999.

MPEG-7 Requirements Group. *MPEG-7: Context, Objectives and Technical Roadmap, V.12.*, *July 1999 ISO/IEC SC29/WG11 N2861.* Vancouver 1999.

MPEG7 Requirements Group. *Results of MPEG-7 Technology Proposal Evaluations and Recommendations.*, *MPEG Seoul Meeting, March 1999 ISO/IEC SC29/WG11 N2730.* Vancouver 1999.

MPEG-7 Video Group. *MPEG-7 Visual Part of the eXperimentation Model Version 2.0.*, *ISO/IEC SC29/WG11 N2822.* Vancouver 1999.

Myaeng, S., D. Shin, W. Chun, and K. Lee. “MIRAGE: A prototype for a multimedia digital library.” *Journal of the Korean Information Science Society* 3, no. 6, December (1997): 605-618.

Owen, C., T. Pearson, and S. Arnold. “Meeting the Challenge of Film Research in the Electronic Age.” *D-Lib Magazine* 6, no. 3, March (2000). www.dlib.org/dlib/march00/owen/03owen.html, (acc. Oct 2000).

PRESTO. *Key Links Systems Specification Document*. 2001. http://presto.joanneum.ac.at/Public/D32.pdf. (acc. March 2002).

PRESTO. *Existing and Emerging Technologies*. 2001. http://presto.joanneum.ac.at/Public/D31.pdf. (acc. March 2002).

PRESTO. *Archiving Preservation and Exploitation Requirements.* 2001. http://presto.joanneum.ac.at/Public/D2.pdf. (acc. March 2002),

Wactlar, H. D. “Informedia Digital Video Library.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/07wactlar.html, (acc. Oct 2000).

Weise, J. “Overview of Media Database Issues and Technologies.” *OIT Digital Media Solutions,* www.oit.itd.umich.edu/projects/DMI/answers/pickmediadb.html, (acc. Jan 2000).

Yeo, B. L., and M. M. Yeung. “Classification, simplification, and dynamic visualization of scene transition graphs for video browsing.” In *Storage and Retrieval for Image and Video Databases VI,*, edited by I K Sethi and R C Jain, 1998. 60-70.

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

7. Preservation

"A Draft Research Agenda for the Preservation of Physical Format Digital Publicatons." National Library of Australia, www.nla.gov.au/policy/rsagenda.html, (acc. Jan 2000), (created Dec 1999*).

"Computer Conservation Society." www.cs.man.ac.uk/CCS, (acc. Oct 2000).

"Conservation Implications of Digitization Projects." 1999, Library of Congress, memory.loc.gov/ammem/techdocs/conserv83199a.pdf, (acc. Oct 2000), (created Sept 1999).

"Ein neues Konvertierungsprogramm der SUB Göttingen/GDZ zur Umwandlung bibliographischer Pica/GBV-Daten in RDF/XML." Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/gdz-tools/pica2xml/pica2xml_index_de.html, (acc. Oct 2000), (created 1997).

"Functional Requirements for Evidence in Recordkeeping." University of Pittsburgh, School of Information Science, www.lis.pitt.edu/~nhprc, (acc. Oct 2000).

"Guidance on expressing the Dublin Core within the Resource Description Framework (RDF)." Dublin Core Metadata Initiative, www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/, (acc. Oct 2000), (created July 1999*).

"ISO Archiving Standards - Reference Model Papers." ssdoo.gsfc.nasa.gov/nost/isoas/, (acc. Oct 2000), (created July 2000*).

"KODAK PHOTO CD Technology/Transferring Images." *Kodak Service & Support Frequently Asked Questions,* www.kodak.com/cluster/global/en/service/faqs/faq1001c.shtml, (acc. Oct 2000), (created Nov 1998*).

"NEDLIB - National Electronic Deposit Library - EU-funded project to develop a European library of digital documents in a stable archival environment." www.kb.nl/coop/nedlib/, (acc. Aug 2000), (created Jan 1998).

"RLG Working Group on Preservation Issues of Metadata: Final Report." Research Libraries Group, www.rlg.org/preserv/presmeta.html, (acc. Oct 2000), (created May 1998).

"RLG-DLF Task Force on Policy & Practice for Long-term Retention of Digital Materials." Research Libraries Group, www.rlg.org/preserv/digrlgdlf99.html, (acc. Oct 2000), (created July 1999*).

"SAA: The Preservation of Digitized Reproductions." 2000, The Society of American Archivists, Chicago, IL, http://www.archivists.org/statements/digitize.html, (acc. Oct 2000), (created June 1997).

"The Conversion Process." Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html#conversion, (acc. Oct 2000), (created June 1999).

"The Photographic Materials Group of the American Institute for the Conservation of Historic & Artistic Works." Boston Art Conservation, www.bosartconserv.com/pmg.htm, (acc. Oct 2000).

Abid, A. "Memory of the World: Preserving our Documentary Heritage." 1997, Information and Informatics

Division. UNESCO, www.unesco.org/webworld/memory/Abid.htm, (acc. Oct 2000), (created July 1997).

Anderson, U. *Workshop on Electronic Archiving: An Evaluation of the Sesam Report.* Stockholm: Swedish National Archives, Astra AB, 1997.

Arms, C. R. “Keeping Memory Alive: Practices for Preserving Digital: Content at the National Digital Library Program of the Library of Congress.” *RLG DigiNews* 4, no. 3, June 15 (2000). www.rlg.org/preserv/diginews/diginews4-3.html#feature1, (acc. Oct 2000).

Ashley, K. “Digital archive costs: facts and fallacies.” In *Proceedings of the DLM-Forum on Electronic Records - European citizens and electronic information: the memory of the information society, Brussels, 18-19 October 1999, INSAR: European Archives News, Supplement IV*, Luxembourg: Office for Official Publications of the European Communities, 2000. 121-128.

Ashley, K. “NDAD: the UK National Digital Archive of Datasets.” In *Electronic access: archives in the new millennium: proceedings, 3-4 June 1999*, London: Public Record Office, 1999. 49-51.

Atkinson, R. “Selection for Preservation: A Materialistic Approach.” *Library Resources and Technical Services* 30 (1986): 34-62.

Banks, P. N., and R. Pilette, (eds.) *Preservation: issues and planning*. 2000, Chicago, IL: American Library Association.

Barry, R. E. “Electronic Document and Records Management Systems: Towards a Methodology for Requirements Definitions.” *Information Management and Technology* 27, no. 6, November (1994): 251-256.

Bearman, D. “An indefensible bastion: archives as repository in the electronic age.” In *Archival management of electronic records. Archives and Museum Informatics Technical Report No. 13*, edited by D Bearman, Pittsburgh, PA: Archives and Museum Informatics, 1991. 14-24.

Bearman, D. “Diplomatics, Weberian bureaucracy, and the management of electronic records in Europe and America.” *American Archivist* 55 (1992): 161-181.

Bearman, D. “Reality and Chimeras in the Preservation of Digital Records.” *D-Lib Magazine* 5, no. 4, April (1999). www.dlib.org/dlib/april99/bearman/04bearman.html, (acc. Oct 2000).

Bearman, D. “The Implications of *Armstrong v. Executive Office of the President* for the Archival Management of Electronic Records.” *American Archivist* 56 (1993): 150-60.

Bearman, D. “Virtual Archives.” *ICA Meeting, September 1996, Beijing,* University of Pittsburgh, www.lis.pitt.edu/~nhprc/prog6.html, (acc. Oct 2000), (created Sept 1996).

Bearman, D. A., and R. H. Lytle. “The Power of the Principle of Provenance.” *Archivaria* 21 (1985-86): 14-27.

Bearman, D. *Archival Methods, Technical Report 3, no.1.* Pittsburgh, PA: Archives and Museum Informatics, 1989.

Bearman, D., and J. Trant. “Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process.” *D-Lib Magazine* 4, no. 6, June (1998). www.dlib.org/dlib/june98/06bearman.html, (acc. Oct 2000).

Bearman, D., and J. Trant. “Electronic Records Research Working Meeting May 28-30, 1997. A Report from the Archives Community.” *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/07bearman.html, (acc. Oct 2000).

Bearman, D., and K. Sochats. “Metadata Requirements for Evidence.” University of Pittsburgh, www.lis.pitt/edu/~nhprc/BACartic.html, (acc. Oct 2000).

Bearman, D., and M. Hedstrom. “Reinventing archives for electronic records: alternative service delivery options.” In *Electronic records management program strategies*, edited by M Hedstrom, Pittsburgh, PA: Archives and Museum Informatics, 1993. 82-98.

Bearman, D., and R. H. Lytle. “The power of the principle of provenance.” *Archivaria* 21 (1985-86): 14-27.

Bearman, D., and W. Duff. “Grounding archival description in the functional requirements for evidence.”

Archivaria 41 (1996): 275-281.

Bearman, v. *Electronic evidence: strategies for managing records in contemporary organizations.* Pittsburgh, PA: Archives and Museum Informatics, 1994.

Bell, L. “The archival implications of machine-readable records.” *Archivum* 26, no. 1 (1979): 85-92.

Bennett, J. C. “A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material.” *eLib: The Electronic Libraries Programme, eLib Supporting Studies,* 1997, (also published as BLRIC Report 50, 1997), www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html, (acc. Oct 2000), (created 1997).

Bennett, J. C. “Archiving and Data Management.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 6-11.

Bewley, R., D. Donaghue, V. Gaffney, M. van Leusen, and A. Wise, (eds.) *Archiving Aerial Photography and Remote Sensing Data: A Guide to Good Practice*. 1999, Oxford: Oxbow Books for the Arts and Humanities Data Service.

Bigourdan, J. L., and J. M. Reilly. *Environment and Enclosures in Film Preservation, Final Report to National Endowment for the Humanities, Grant #PS20802-94.*: Image Permanence Institute, Rochester Institute of Technology, 1997.

Bikson, T. K., and E. J. Frinking. *Preserving the Present: Toward Viable Electronic Records.* Den Haag: Sdu Publishers, 1993.

Bikson, T. K., and S. A. Law. “Electronic Information Media and Records Management Methods: A Survey of Practices in United States Organizations.” *Information Society* 9, no. 2: 125-44.

Braudaway, G. “Restoration of faded photographic transparencies by digital image processing.” In *Proceedings of IS&T’s 46th Annual Conference*, Springfield, VA: IS&T, 1993. 287-289.

Clark, S. *Preservation of Photographic Materials.*: National Preservation Office, 1999.

Coe, B., and M. Haworth-Booth. *A Guide to Early Photographic Processes.* London 1983.

Commission on Preservation and Access, and the Research Libraries Group. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Washington, D.C.: Commission on Preservation and Access, 1996.

Commission on Preservation and Access. *Preserving the Illustrated Text. Report of the Joint Task Force on Text and Image.* Washington, DC: Commission on Preservation and Access, 1992.

Conway, P. “Digitizing preservation.” *Library Journal* December (1994): 42-45.

Conway, P. “Selecting Microfilm for Digital Preservation: A Case Study from Project Open Book.” *Library Resources and Technical Services* 40, no. 1 (1996a): 67-77.

Conway, P. “Yale University Library’s Project Open Book.” *D-Lib Magazine* 2, no. 2, February (1996). www.dlib.org/dlib/february96/yale/02conway.html, (acc. Oct 2000).

Conway, P. *Preservation in the Digital World.* Washington, D.C.: Commission on Preservation and Access, 1996b.

Conway, P. *Selecting Microfilm for Digital Preservation: A Demonstration Project: Performance Report on the production conversion phase of Project Open Book.* New Haven, Conn.: Yale University Library, 1996a.

Cook, T. “Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era.” *Archives and Manuscripts* 22, no. November (1994): 300-328.

Cox, R. J. “Re-Discovering the Archival Mission: The Recordkeeping Functional Requirements Project at the University of Pittsburgh, A Progress Report.” *Archives and Museum Management* 8, no. 4 (1994): 279-300.

Cox, R. J. “The Record in the Information Age: A Progress on Reflection and Research.” *Records & Retrieval Report* 12, no. 1 (1996): 1-16.

Cullen, C. T., P. B. Hirtle, D. Levy, C. A. Lynch, and J. Rothenberg. *Authenticity in a Digital Environment.*: Council on Library and Information Resources, 2000.

Davidson, J., and L. Moscato. “Towards an Electronic Records Management Program: The University of Melbourne.” *Archives and Manuscripts* 22, no. 1 (1994): 124-135.

Day, M. “Issues and Approaches to Preservation Metadata.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 73-84.

Day, M. “Preservation of Electronic Information: A Bibliography.” *UKOLN: The UK Office for Library and Information Networking,* 1997, homes.ukoln.ac.uk/~lismd/preservation.html, (acc. Oct 2000), (created Jan 1997).

Day, M. W. “CEDARS: Digital Preservation and Metadata.” In *Sixth DELOS Workshop: Preservation of Digital Information, Tomar, Portugal 17-19 June 1998. ECRIM-98-W003*, Le Chesnay: European Research Consortium for Informatics and Mathematics, 1998. 53-58.

Dementi, M. R. “Access and Archiving as a New Paradigm.” *The Journal of Electronic Publishing* 3, no. 3, March (1998). www.press.umich.edu/jep/03-03/dementi.html, (acc. Oct 2000).

Dempsey, L., and S. L. Weibel. “The Warwick Metadata Workshop: a framework for the deployment of resource description.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/07weibel.html, (acc. Oct 2000).

Derby, D. “Caring for Your Photographs.” 1997, American Institute for Conservation, palimpsest.stanford.edu/aic/treasure/photos.html, (acc. Oct 2000).

Dollar, C. M. *Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Practices, Information and Documentation.* Edited by Oddo Bucci, *Series 1.* Macerata, Italy: University of Macerata, 1992.

Dollar, C. M. *Authentic Electronic Records: Strategies for Long Term Access.* Chicago, IL: Cohasset Associates, Inc., 1999.

Dowler, L. “The Role of Use in Defining Archival Practice and Principles: A Research Agenda for the Availability and Use of Records.” *American Archivist* 51 (Winter and Spring) (1988): 74-95.

Duchein, M. “Theoretical Principles and Practical Problems of Respect des Fonds.” *Archivaria* 16, Summer (1983): 64-82.

Duff, W. “Ensuring the Preservation of Reliable Evidence: A Research Project Funded by the NHPRC.” *Archivaria* 42 (1995): 28-45.

Duranti, L., and H. MacNeill. “The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project.” *Archivaria* 42 (1995): 46-67.

Eden, P. *A Model for Assessing Preservation Needs in Libraries.*, *British Library Research and Innovation Centre, BLRIC report no. 125* 1998.

Erlandsson, A. “The Principle of Provenance and the Concept of Records Creator and Record: Legal Development.” In *The Principle of Provenance*, Stockholm: Swedish National Archive, 1994. 33-49.

Erlandsson, A. *Electronic Records Management: A Literature Review.*, *International Council on Archives (ICA) Study* 1996. www.ica.org/cgi-bin/ica.pl?0508_e, (acc. Oct 2000).

Feather, J. “The Preservation of Information: Principles and Practice of Format Conversion.” *Library Review* 40, no. 6 (1991): 7-8.

Feather, J. *Preservation and the Management of Library Collections.* London 1991.

Feather, J., G. Matthews, and P. Eden. *Preservation Management. Policies and Practices in British Libraries.* Aldershot/Brookfield 1996.

Feeney, M., (ed.) *Digital culture: maximising the nations investment: a synthesis of JISC/NPO studies on the preservation of electronic materials*. 1999, London: British Library Board.

Foster, S., (ed.) *Proceedings of the First International "Memory of the World" Conference, Oslo, 3-5 June 1996*. 1996, Olso: UNESCO.

Foster, S., R. Russel, J. Lyall, and D. Marchall. *Memory of the World: General Guidelines to Safeguard Documentary Heritage.*, *Report CII-95/WS-11.* Paris: UNESCO, 1995.

Fox, L. *Preservation Microfilming: A Guide for Librarians and Archivists.* 2nd ed. Chicago, IL: American Library Association, 1996.

Fresko, M., and K. Tombs. *Digital Preservation Guidelines: The state of art in libraries, museums and archives.*: European Commission, 1998.

Frey, F. “Digitize to Preserve - Photographic Collections Facing the Next Millennium.” In *Proceedings, IS&T 50th Annual Conference*, Boston, 1997. 713-715.

Frey, F. “How Long Will It Last? - Life Expectancy of Information Media.” 1999, City of Newport News, www.newport-news.va.us/wwdept/infosite/tsld001.htm, (acc. Oct 2000), (created Oct 1999).

Frey, F., R. Gschwind, and L. Rosenthaler. “Electronic Imaging, a Tool for the Reconstruction of Faded Color Photographs and Motion Pictures.” In *Proceedings, IS&T Fourth Color Imaging Conference*, 1996. 39-44.

Garrett, J. R. “Task Force on Archiving of Digital Information.” *D-Lib Magazine* 1, no. 9, September (1995). www.dlib.org/dlib/september95/09garrett.html, (acc. Oct 2000).

Gilliland-Swetland, A. J. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment.*: Council on Library and Information Resources, 2000.

Gould, S., and R. Ebdon, (ed.). *IFLA/UNESCO Survey on Digitisation and Preservation.*, *International Preservation 2* 1999.

Gracy, D. “Archives and Society: The First Archival Revolution.” *American Archivist* 47 (Winter) (1984): 7-10.

Graham, P. S. “Long-term Intellectual Preservation.” In *Digital Imaging Technology for Preservation*, edited by N E Elkington, Mountain View, CA: Research Libraries Group, 1994. 41-57.

Green, A., J. Dionne, and M. Dennis. *Preserving the Whole: A Two-Track Approach to Rescuing Social Science Data and Metadata.* Washington, DC: Council on Library and Information Resources, 1999. http://www.clir.org/pubs/reports/pub83/pub83.pdf, (acc. Oct 2000).

Greenstein, D., and N. Beagrie. “Managing Digital Collections: AHDS Policies, Standards, and Practices.” Arts and Humanities Data Service, www.ahds.ac.uk/public/srg.html, (acc. Oct 2000), (created Dec 1998).

Gulbins, J., M. Seyfried, and H. Strack-Zimmermann. *Elektronische Archivierungssysteme.* Berlin: Heidelberg & others, 1993.

Haynes, D. S., T. Jowett, and M. Blake. *Responsibiliy for digital archiving and long term access to digital data.*, *Elib study P1.* London: JISC and NPO, 1997.

Hedstrom, M. “Principles for Creating a Basic Preservation Strategey.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 140.

Hedstrom, M. “The Role of National Initiatives in Digital Preservation.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 85-87.

Hedstrom, M. “Understanding Electronic Incunabula: A Framework for Research on Electronic Records.” *American Archivist* 54, no. 3, (summer) (1991): 334-354.

Heminger, A. R., and S. B. Roberston. “Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents.” In *Sixth DELOS Workshop: Preservation of Digital Information. ECRIM-98-W003*, Le Chesnay: European Research Consortium for Informatics and Mathematics, 1998. 35-43.

Hendley, T. *Comparison of Methods & Costs of Digital Preservation: A Consultancy Study.*, *British Library Research and Innovation Report 106.* London: British Library Research and Innovation Centre, 1998.

Hendley, T. *The Preservation of Digital Material.* London: British Library Research and Development Department, 1996.

Hodge, G. M. “Best Practices for Digital Archiving: An Information Life Cycle Approach.” *D-Lib Magazine* 6, no. 1, January (2000). www.dlib.org/dlib/january00/01hodge.html, (acc. Oct 2000).

Indiana University Electronic Records Project. “A Methodology for Evaluating Recordkeeping Systems.” *Draft (March),* 1997, Indiana University, www.indiana.edu/~libarche/97method.html, (acc. Oct 2000), (created April 1998*).

Jürgens, M. C. “Preservation of Ink Jet Hardcopies.” Royal Netherlands Academy of Arts and Sciences, www.knaw.nl/ecpa/publ/jurgens.html, (acc. Oct 2000), (created Aug 1999).

Kahle, B. “Preserving the Internet.” *Scientific American* 276, no. 3, March (1997). www.sciam.com/0397issue/0397kahle.html, (acc. Oct 2000).

Kenney, A. R., and L. K. Personius. *Joint Study in Digital Preservation, Report: Phase I, Digital Capture, Paper Facsimiles, and Network Access.* Washington, DC: Commission on Preservation and Access, 1992. www.clir.org/pubs/reports/joint/index.html, (acc. Oct 2000).

Kenney, A. R., and P. Conway. “From Analog to Digital: Extending the Preservation Toolkit.” In *Digital Imaging Technology for Preservation*, edited by N E Elkington, Mountain View, CA: Research Libraries Group, 1994. 11-24.

Ketelaar, E. “Exploitation of New Archival Materials.” In *Proceedings of the 11th International Congress on Archives*, Munich: K. G. Sauer, 1989. 189-99.

Kiernan, K. “Digital Preservation, restoration, and the dissemination of medieval manuscripts.” In *Gateways, gatekeepers, and roles in the information omniverse: proceedings of the third symposium: November 13-15, 1993*, edited by A Okerson and D Mogge, Washington, DC, 1994.

Kirtley, T. “Preservation and digitisation: principles, practice and policies: The 1996 National Preservation Office Conference.” *Library Conservation News* 53, no. Winter (1997).

Klijn, E., and Y. de Lusenet. *In the Picture: Preservation and Digitisation of European Photographic Collections.* Amsterdam: European Commission on Preservation and Access, 2000.

Kuny, T. “The Digital Dark Ages? Challenges in the Preservation of Electronic Information.” *International Preservation News* 17, no. May (1998): 8-13.

Lavender, K., P. Conway, S. J. Dalton, J. F. Dean, and J. Francis. *Managing preservation technologies for the new century*. Chicago IL: American Library Association, 2000. Audio Cassette.

Lesk, M. “Preserving Digital Objects: Recurrent Needs and Challenges.” 1995, www.lesk.com/mlesk/auspres/aus.html, (acc. Oct 2000).

Lesk, M. *Preservation of New Technology. A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access.* Washington, DC: Commission on Preservation and Access, 1995.

Levy, D. “Heroic Measure: Reflections on the Possibility and Purpose of Digital Preservation.” In *ACM Digital Library, International Conference on Digital Libraries: Proceedings of the Third ACM Conference on Digital Libraries, June 21-23, 1998, Pittsburgh, PA*, 1998.

Library of Congress: National Digital Library Program and the Conservation Division. “Conservation Implications of Digitization Projects.” Library of Congress, American Memory, memory.loc.gov/ammem/techdocs/conservation.html, (acc. Oct 2000), (created Sept 1999).

Lyman, P., and B. Kahle. “Archiving Digital Cultural Artifacts: Organizing an Agenda for Action.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/07lyman.html, (acc. Oct 2000).

Lynch, C. “Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information.” *D-Lib Magazine* 5, no. 9, September (1999). www.dlib.org/dlib/september99/09lynch.html, (acc. Oct 2000).

Manes, S. “Time and Technology Threaten Digital Archives.” *New York Times* April 7 (1998): f-4.

McIntyre, J. E. “Protecting the Physical Form.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 33-38.

Mohlenrich, J., (ed.) *Preservation of electronic formats & electronic formats for preservation*. 1993, Atkinson, Wis.: Highsmith.

Moore, R., C. Baru, A. Rajasekar, B. Ludaescher, R. Marciano, M. Wan, W. Schroeder, and A. Gupta. “Collection-Based Persistent Digital Archives - Part 1.” *D-Lib Magazine* 6, no. 3, March (2000). www.dlib.org/dlib/march00/moore/03moore-pt1.html, (acc. Oct 2000).

Moore, R., C. Baru, A. Rajasekar, B. Ludaescher, R. Marciano, M. Wan, W. Schroeder, and A. Gupta. “Collection-Based Persistent Digital Archives - Part 2.” *D-Lib Magazine* 6, no. 4, April (2000). www.dlib.org/dlib/april00/moore/04moore-pt2.html, (acc. Oct 2000).

Mowat, I. K. M. “Art Objects or Working Tools: The Present and Future Funding of Preservation.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 81-84.

Museums and Galleries Commission. *Care of Photographic Materials and Magnetic Media.* London: Museums and Galleries Commission, 1998.

National Archives and Records Administration (NARA). *Digital Imaging and Optical Digital Data Disk Storage Systems: Long-Term Access Strategies for Federal Agencies.*, *Technical Information Paper no. 12*: NARA, 1994.

National Archives and Records Administration (NARA). *Management, Preservation and Access for Electronic Records with Enduring Value.* Washington, DC: NARA, 1991.

National Archives and Records Administration Electronic Records Information/GRS20. “Electronic Records Work Group.” NARA, www.nara.gov/records/grs20/index.html, (acc. Oct 2000), (created Nov 1998).

National Archives of Australia, *Archiving Web Resources: policy and guidelines for keeping records of web-based activity in the Commonwealth Government*, National Archives of Australia, 2001, http://www.naa.gov.au/recordkeeping/er/web_records/intro.html.

National Library of Australia. “PADI: Preserving Access to Digital Information.” www.nla.gov.au/padi, (acc. Oct 2000).

National Research Council. *Preserving Scientific Data on our Physical Universe: A New Strategy for Archiving the Nation’s Scientific Information Resources.* Washington, DC: National Academy Press, 1995.

National Research Council. *Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers.* Washington, DC: National Academy Press, 1995.

Nurminen, A. “Recording, Processing, and Archiving Carbonised Papyri.” 1995, Helsinki University of Technology, www.cs.hut.fi/papyrus/TOC.html, (acc. Oct 2000).

O’Toole, J. M. “On the Idea of Permanence.” *American Archivist* 52, no. 1 (1989): 10-25.

Ogden, B. “The Preservation Perspective.” In *Selecting Library and Archive Collections for Digital Reformatting: Proceedings from an RLG Symposium held November 5-6, 1995, in Washington, DC*, edited by Ricky Erway, Mountain View, CA: RLG, 1996.

Ogden, S. “Temperature, Relative Humidity, Light and Air Quality. Basic Guidelines for Preservation.” 1999, Northeast Document Conservation Center (NEDCC), www.nedcc.org/plam3/tleaf21.htm, (acc. Oct 2000).

Ogden, S., (ed.) *Preservation of Library and Archival Materials: a Manual*. 3rd edition, revised and expanded ed. 1999.

Perlman, E., and I. Kallen. “Common Internet File Formats.” 1995, www.matisse.net/files/formats.html, (acc. Oct 2000), (created Dec 1995*).

Phillips, M. E. “Ensuring Long-Term Access to Online Publications.” *JEP: The Journal of Electronic Publishing* 4,

June (1999). www.press.umich.edu/jep/04-04/phillips.html, (acc. Oct 2000).

Picture Elements, I. *Guidelines for Electronic Preservation of Visual Materials.* revisions 1.1, 2 ed., *Report to the Library of Congress, Preservation Directorate* 1995.

Puglia, S. “Creating Permanent and Durable Information, Physical Media and Storage Standards.” *CRM* 2 (1999). tps.cr.nps.gov/crm/archive/22-2/22-02-10.pdf, (acc. Oct 2000).

Reilly, B. F., J. M. Reilly, D. H. Norris, S. T. Puglia, P. De Stefano, J. Van Haaften, and P. McClung. *Photograph Preservation and the Research Library.* Edited by Jennifer Porro. Mountain View, CA: The Research Libraries Group, 1991.

Reilly, J. M. *Care and Identification of 19th Century Photographic Prints.*: Eastman Kodak Company, 1986.

Reilly, J. M., D. W. Nishimura, and E. Zinn. *New Tools for Preservation: Assessing Long-Term Environmental Effects on Library and Archives Collections.* Washington, DC: Commission on Preservation and Access, 1995.

Rhinehart, R. “The Straw that Broke the Museum’s Back? Collecting and Preserving Digital Media Art Works for the Next Century.” *Switch: Journal of New Media Art* v6n1.switch.sjsu.edu/web/v6n1/articlea.htm, (acc. Oct 2000).

Rieger, O. Y. “Projects to Programs: Developing A Digital Preservation Policy.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 135-152.

RLG Working Group. “Preservation Issues of Metadata.” *Final Report,* 1998, RLG, www.rlg.org/preserv/presmeta.html, (acc. Oct 2000), (created May 1998).

RLG. “PRESERV - The RLG Preservation Program.” Research Libraries Group, www.rlg.org/preserv, (acc. Oct 2000), (created Sept 2000*).

Roberts, H. E. ""Second Hand Images: The Roles of Surrogates in Artistic and Cultural Exchange". Paper delivered as the Keynote Address for the Visual Resources Satellite Session of the XXVIIIth International Congress of the History of Art (CIHA), Berlin, July 1992.” *Visual Resources* 9, no. 4 (1993): 335-346.

Robertson, S. B. “Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents.” MSc, Air Force Institute of Technology, 1996.

Ross, S. “Changing trains at Wigan: Digital preservation and the future of scholarship.” *The JISC/NPO Digital Preservation workshop, 3/4 March 1999, Scarman House, University of Warwick,* 1999, www.leeds.ac.uk/cedars/OTHER/SRoss.htm, (acc. Oct 2000), (created March 1998).

Ross, S. “Preserving and Maintaining Electronic Resources in the Visual Arts for the Next Century.” *Computers and the History of Art* 7, no. 2 (1997): 35-50.

Ross, S. “The Expanding World of Electronic Information and the Past’s Future.” In *Historians and Electronic Artefacts*, edited by E Higgs, London, 1997. 6-28.

Ross, S., and A. Gow. “Digital Archaeology: Rescuing Neglected and Damaged Data Resources.” 1999, eLib: The Electronic Libraries Programme, eLib Supporting Studies, www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf, (acc. Oct 2000), (created Feb 1999).

Rothenberg, J. “Ensuring the Longevity of Digital Documents.” *Scientific American* 272, no. 1, January (1995): 24-9.

Rothenberg, J. “Metadata to Support Data Quality and Longevity.” *Proceedings of the 1st IEEE Metadata Conference, NOAA Complex, Silver Spring, Md., 16-18 April,* 1996, www.computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html, (acc. Oct 2000).

Rothenberg, J. “Survival of the Digits.” *Communicating Business* 4, no. Spring (1995).

Rothenberg, J. “Using Emulation to Preserve Digital Information.” *(paper presented at NSF Workshop on Data Archival and Information Presentation, March 26-27, 1999),* MU Computer Engineering, Computer Science, cecssrv1.cecs.missouri.edu/NSFWorkshop/ppaper3.html, (acc. Oct 2000), (created March 1999).

Rothenberg, J. *Avoiding Technical Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Washington, DC: Council on Library and Information Resources (CLIR), 1999. www.clir.org/pubs/reports/rothenberg/contents.html, (acc. Oct 2000).

Rothenberg, J. *Digital Information Lasts Forever - Or Five Years, Whichever Comes First*: Rand, V-079, 1997. Video.

Russell, K. “Digital Preservation: Ensuring Access to Digital Materials Into the Future.” CEDARS, www.leeds.ac.uk/cedars/Chapter.htm, (acc. Oct 2000), (created July 1999).

Russell, K., and E. Weinberger. “Cost elements of digital preservation. Draft of 31 May 2000.” www.leeds.ac.uk/cedars/documents/CIW01r.html, (acc. Oct 2000), (created May 2000).

Sawyer, D., and L. Reich. “Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-W-5.0.” *White Book*, no. 5 (1999).

Scherlis, W. L. “Repository Interoperability Workshop: Towards a Repository Reference Model.” *D-Lib Magazine* 2, no. 10, October (1996). www.dlib.org/dlib/october96/workshop/10scherlis.html, (acc. Oct 2000).

Smith, A. “Preservation in the Future Tense.” *CLIR Issues* 3, no. May/June (1998). www.clir.org/pubs/issues/issues03.html#preserve, (acc. Oct 2000).

Smith, A. *The Future of the Past: Preservation in American Research Libraries.* Washington, DC: Council on Library and Information Resources (CLIR), 1999. www.clir.org/cpa/reports/pub82/pub82text.html, (acc. Oct 2000).

Society of American Archivists. “Statement on the Preservation of Digitized Reproductions.” 1997, www.archivists.org/governance/handbook/app_k6.htm, (acc. Oct 2000), (created June 1997).

Stepanek, M. “From Digits to Dust.” *Businessweek* 3574, no. April 20 (1998): 128.

Stone, A., and M. Day. “Cedars Preservation Metadata Elements.” *Cedars Project Document AIW02,* The Cedars Project, users.ox.ac.uk/~cedars/Papers/AIW02.html, (acc. Oct 2000), (created Feb 1999).

Swora, T. “The Preservation Context.” In *Selecting Library and Archive Collections for Digital Reformatting: Proceedings from an RLG Symposium held November 5-6, 1995, in Washington, DC*, edited by Ricky Erway, Mountain View, CA: RLG, 1996.

UKOLN: the UK Office for Library and Information Networking. “Long term preservation of electronic materials: a JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib) organised by UKOLN, 27th and 28th November 1995 at the University of Warwick.” *British Library R&D Report 6238,* 1995, www.ukoln.ac.uk/services/elib/papers/other/preservation/intro.html, (acc. Oct 2000), (created Nov 1995).

Van Bogart, J. W. C. “Long-Term Preservation of Digital Materials.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 17-36.

van der Werf-Davelaar, T. “Long-term Preservation of Electronic Publications: The NEDLIB project.” *D-Lib Magazine* 5, no. 9, September (1999). www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html, (acc. Oct 2000).

Vavra, J. K. “Confronting Obselescence: An Archival Lifestyle.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 138-139.

Waters, D. J. “Digital Archiving: The Report of the CPA/RLG Task Force.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 37-48.

Weber, H., and M. Dörr. *Digitisation as a method of preservation? Final report of a working group of the Deutsche Forschungsgemeinschaft (German Research Assocation).* Amsterdam: European Commission on Preservation and Access, 1997.

Weibel, S., and E. Miller. “Image Description on the Internet.” *D-Lib Magazine* 3, no. 1, January (1997). www.dlib.org/dlib/january97/oclc/01weibel.html, (acc. Oct 2000).

Wilhelm, H., and C. Brower. *The Permanence and Care of Color Photographs: Traditional and Digital Color Prints, Color Negatives, Slides, and Motion Pictures.* Grinnell, Iowa: Preservation Publishing Company, 1993.

Willis, D. *A Hybrid Systems Approach to Preservation of Printed Materials.* Washington, DC: Commission on Preservation and Access, 1992.

Yoder, S. M. “Sustainability through Integration.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 173-174.

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

8. Quality Assurance (content; metadata)

"Fully Utilizing Photo CD Images: Typical Monitor Luminance for Photo YCC Luma Values - Article Number 6." Eastman Kodak Company, www.kodak.com/US/en/digital/techInfo/pcd-102.shtml, (acc. Jan 2000), (created Aug 1995).

"Image Quality Evaluation." www.mitre.org/centers/cafc3/mtf/index.html, (acc. Oct 2000), (created June 2000*).

"Instrumental Quality Control." In *Fundamentals of Color and Appearance*, New Windsor, NY: MacBeth, 1996.

"Monitor Adjustment Target." National Archives and Records Administration, NARA Archival Information Locator (NAIL), www.nara.gov/nara/target.html, (acc. Jan 2000), (created March 2000*).

"Munsell Color Vision Tests." GretagMacbeth, www.munsell.com/munvis.htm, (acc. Jan 2000).

"Preservation Resources, Scanning Questionnaire." *Quality Control Documentation,* Online Computer Library Center, Preservation Resources, www.oclc.org/presres/scanning/qa.htm, (acc. Aug 2000).

"Scanners and Digital Cameras." *RLG DigiNews* 1, no. 1, April 15 (1997). www.thames.rlg.org/preserv/diginews/diginews1.html, (acc. Oct 2000).

"Scanners." Technical Advisory Service for Images (TASI), www.tasi.ac.uk/building/scannersprint.html, (acc. Oct 2000).

"Setting the Resolution." *Photography Collection,* Denver Public Library, Western History/Genealogy Department, gowest.coalliance.org/calib.htm, (acc. Oct 2000).

"Tests." Center for Digital Imaging Inc., www.cdiny.com/tests.html, (acc. Oct 2000), (created *forthcoming*).

"The Photographic Materials Group of the American Institute for the Conservation of Historic & Artistic Works." Boston Art Conservation, www.bosartconserv.com/pmg.htm, (acc. Oct 2000).

"Workflow Guidelines: 2. Pre-scanning Preparation and Post-scanning Processing of the Originals." TASI, www.tasi.ac.uk/framework/collections/jidi-workflow.html#wg2, (acc. Oct 2000).

AGFA. *An Introduction to Digital Scanning.* Belgium: Agfa-Gevaert, 1994.

ANSI/AIIM MS44-1988. *Recommended Practice for Quality Control of Image Scanners.* Silver Spring, MD: Association for Information and Image Management, 1988.

Association for Information and Image Management. *Practice for Operational Procedures/Inspection and Quality Control of First-Generation, Silver Microfilm of Documents.* Silver Spring, MD: Association for Information and Image Management, 1998.

Avedon, D. *Quality Control of Electronic Images.* Silver Spring, MD: Association of Information and Image Management, 1997.

Barten, P. “MTF, CSF, and SQRI for Image Quality Analysis, short course notes.” In *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology*, 1996.

Braunegg, D. J., R. D. Forkert, and N. B. Nill. *Rescaling Digital Fingerprints: Techniques and Image Quality Effects, MITRE Report for FBI Fingerprint Project.* 1995.

Burns, P. D. “What Does the Colour Characterisation of Digital Still Cameras Standard Promise for the Digital Imaging Community?” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 77.

Chambliss, M. A., J. A. Dawson, and E. J. Borg. “Measuring the MTF of Undersampled IRFA Sensors Using 2D Discrete Transformations.” In *SPIE Proceedings: Infrared Imaging Systems, Design, Analysis, Modeling, and Testing VI*, edited by G C Holst, 1995. 312-324.

CIE Technical Committee 8-02. “Colour Difference Evaluation in Images.” International Commission on Illumination, www.colour.org/tc8-02, (acc. Oct 2000), (created Oct 1998).

Engeldrum, P. “A Framework for Image Quality Models.” *Journal of Imaging Science and Technology* 39, no. 4 (1995): 312-319.

Engeldrum, P. “Introduction to Image Quality.” In *IS&T 11th International Congress on Advances in Non-Impact Printing Technologies, November 1995*, 1995.

Ester, M. “Image Quality and Viewer Perception.” *Leonardo* 23, no. 1 (1990): 51-63.

Ester, M. “Image Use in Art - Historical Practice.” In *Gateways, Gatekeeers, and Roles in the Information Omniverse, Proceedings form the Third Symposium*, edited by Ann Okerson and Dru Mogge, Washingon, DC: Association of Research Libraries - Association of American University Presses, 1993.

Forkert, R. D., and et al. *Test Procedures for Verifying IAFIS Scanner Image Quality Requirements, MITRE Report for FBI Fingerprint Project.* 1994.

Fraser, B. “Out of Gamut: Stop the sRGB Bandwagon.” *MacWeek*, no. July 31 (1998). macweek.zdnet.com/1229/op_fraser.html, (acc. Oct 2000).

Frey, F. “Benchmarking Digitizing Systems for Photographs.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 36.

Frey, F., and S. Süsstrunk. “Color Issues to Consider in Pictorial Image Databases.” In *Proceedings, IS&T Fifth Color Imaging Conference*, Scottsdale, AZ, 1997. 112-115.

Frey, F., and S. Süsstrunk. “Image Quality Issues for the Digitization of Photographic Collections.” In *Proceedings, IS&T 49th Annual Conference*, Minneapolis, MN, 1996. 349-353.

Frey, F., R. Gschwind, and L. Rosenthaler. “Electronic Imaging, a Tool for the Reconstruction of Faded Color Photographs and Motion Pictures.” In *Proceedings, IS&T Fourth Color Imaging Conference*, 1996. 39-44.

Gann, R. G. *Desktop Scanners: Image Quality Evaluation.* Upper Saddle River, NJ: Prentice-Hall, 1999.

Gann, R., J. Goode, and R. Wadas. *Reviewing and Testing Desktop Scanners.* Second ed. Greeley, CO: Hewlett Packard Company, 1995.

Gilman, P. B., E. McInerney, and R. Vogel. “Some Simple Techniques for Improving the Print Quality of Digital Images from Kodak Electronic Cameras.” In *International Conference on Digital Printing Technologies, NIP 12*, Springfield, VA: IS&T, 1996. 227-232.

Giorgianni, E. J., and T. E. Madden. *Digital Color Management.* Reading, MA: Addison-Wesley, 1997.

Gladney, H. M., and J. B. Lotspiech. “Safeguarding Digital Library Contents and Users: Assuring Convenient Security and Data Quality.” *D-Lib Magazine* 3, no. 5, May (1997). www.dlib.org/dlib/may97/ibm/05gladney.html, (acc. Oct 2000).

GretagMacbeth. *Fundamentals of Color and Appearance.* New Windsor, NY: GretagMacbeth, 1997.

Grotta, D., and S. W. Grotta. “Image Quality Tests: Digital Cameras.” *ZDNet Products*, no. February 10 (1998). www.zdnet.com/products/content/pcmg/1703/268566.html, (acc. Oct 2000).

Holm, J., and S. Süsstrunk. “An EIQ-Subjective Image Quality Correlation Study.” In *Proceedings IS&T 47th Annual Conference*, Rochester, NY, 1994. 634-640.

Hüfner, J. “Steigerung der Erkennungsgenauigkeit durch maschinellen Abgleich verschiedener durch OCR erzeugter Volltexte (Increasing the Quality of OCR-generated Full Texts by Automatic Comparison).” *Nachrichten für Dokumentation* 48, no. 2 (1997): 79-85.

ISO 12233/FDIS, ISO/TC42, Photography - Electronic Still Picture Cameras - Resolution Measurements. 1999.

ISO 14524/FDIS, ISO/TC42, Photography - Electronic Still Picture Cameras - Methods for Measuring Opto-Electronic Conversion Functions (OECFs). 1999.

ISO 15739/CD, ISO/TC42, Photography - Electronic Still Picture Cameras - Noise Measurements. 1999.

ISO 16067/WD, ISO/TC42, Electronic scanners for photographic images - Spatial resolution measurements: Part 1, Scanners for reflective media. 1999.

ISO 17321/WD, ISOTC42, Graphic Technology and Photography - Colour Characterisation of Electronic Still Cameras (DSCs) Using Colour Targets and Spectral Illumination. 1999.

ISO/DIS 12234-2, Photography - Electronic Still Pictures Cameras - Removable Memory - Part 2: Image Data Format - TIFF/EP. 1998.

ISO/FDIS 12233 Photography, Electronic Still Picture Cameras, Resolution Measurements.

Kenney, A. R., and S. Chapman. *Tutorial. Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality.* Washington, DC: Commission on Preservation and Access, 1995.

Kinoshita, H., and T. Yamamuro. “Effects of Repetitive JPEG Compressions with DCT Block Rearrangement Operation on Image Quality.” *Journal of Imaging Science and Technology* 39, no. 6, November/December (1995): 546-558.

Kipman, Y. “Image Quality Metrics for Digital Image Capture Devices.” In *Proceedings, IS&T 48th Annual Conference*, Washington, DC, 1995. 456-459.

Library of Congress: National Digital Library Program. “Quality Review of Document Images, Internal Training Guide.” *NDLP Internal Documentation,* 1996, memory.loc.gov/ammem/award/docs/docimqr.html, (acc. Oct 2000), (created Sept 1996).

Margulis, D. “How Color Management Failed.” *Electronic Publishing* July (1999).

Marisa, R. “Metadata Quality Control.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 62.

Mayer, M. “Scanners.” *Message posted to Conservation DistList, 9 December 1997,* 1997, palimpsest.stanford.edu/byform/mailing-lists/cdl/1997/1549.html, (acc. Oct 2000), (created Sept 1997).

Means, L. “Digitisation as Transformation and Some Implications for the Arts.” *Leonardo* 17, no. 3 (1984): 195-199.

Moss, J., and R. Dale, (eds.). “Guides to Quality in Visual Resource Imaging Home Page.” 2000, Research Libraries Group, Mountain View, CA, http://www.rlg.org/visguides/index.html, (acc. Oct 2000), (created July 2000).

Nartker, T. A., and S. V. Rice. “OCR Accuracy: UNLV’s Third Annual Test.” *INFORM* 8, no. 8 (1994): 30-36.

Nartker, T. A., S. V. Rice, and J. Kanai. “OCR Accuracy: UNLV’s Second Annual Test.” *INFORM* 8, no. 1 (1994): 40-45.

Nill, N. B., and B. R. Paine. *A Computer Program to Determine the Sine Wave Modulation Transfer Function of Image Scanners, MITRE Report for FBI Fingerprint Project.* 1994.

Poe, R. “Aesthetic Considerations in Tone and Color Management.” In *Proceedings, IS&T Third Color Imaging*

Conference, Scotsdale, AZ, 1995. 164-168.

Poynton, C. “Frequently-Asked Questions About Gamma.” www.inforamp.net/~poynton/GammaFAQ.html, (acc. Oct 2000), (created Aug 1998).

Poynton, C. “Poynton’s Color FAQs.” www.inforamp.net/~poynton/ColorFAQ.html, (acc. Oct 2000), (created April 2000).

Reichenbach, S. E., S. K. Park, and R. Naraynswamy. “Characterizing Digital Image Acquisition Devices.” *Optical Engineering* 30, no. 2 (1991): 170-177.

Reilly, J. M., and F. S. Frey. “Recommendations for the Evaluation of Digital Images Produced from Photographic, Microphotographic, and Various Paper Formats.” 1996, Library of Congress, American Memory, memory.loc.gov/ammem/ipirpt.html, (acc. Oct 2000), (created May 1996).

Rieger, R., and G. Gay. “Tools and Techniques in Evaluating Digital Imaging Projects.” *RLG DigiNews* 3, no. 3, June 15 (1999). www.rlg.org/preserv/diginews/diginews3-3.html#technical1, (acc. Oct 2000).

Rusch-Feja, D., and U. Siebeky. “Evaluation of Usage and Acceptance of Electronic Journals: Results of an Electronic Survey of Max Planck Society Researchers including Usage Statistics from Elsevier, Springer and Academic Press.” *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/rusch-feja/10rusch-feja-summary.html, (acc. Oct 2000).

Shaw, R., (ed.) *Selected Readings in Image Evaluation*. 1976, Washington, DC: Society of Photographic Scientists and Engineers.

Stokes, M., M. Anderson, S. Chandrasekar, and R. Motta. “A Standard Default Color Space for the Internet - sRGB.” *Version 1.10,* 1996, www.w3.org/Graphics/Color/sRGB, (acc. Oct 2000), (created Nov 1996).

The Secrets of Color Management. Randolph, MA: Agfa Educational Publishing, 1997.

UNESCO. “Digitizing Library and Archives Collections.” WebWorld, unesco.org/webworld/highlights/digitisation_0199.html, (acc. Oct 2000).

US Army Corps of Engineers. “Economic Analysis, Executive Summary Report, Optical Disk Imaging.” United States Army Corp of Engineers, http://144.3.144.209/corpusdata/usace/inet/usace-docs/s-r/etl1110-1-168/entire.pdf, (acc. Oct 2000), (created Oct 1997).

Wasner, R. *Display Subsystems for Production Imaging.* Hingham, MA: Rheinner, 1998.

Williams, D. “An Overview of Image-Quality Metrics.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 72-73.

Williams, D. “Benchmarking of rhe ISO 12233 Slanted-edge Spatial Frequency Response Plug-in.” In *Proceedings, IS&T 1998 PICS Conference*, Portland, OR, 1998. 133-137.

Williams, D. “Image Quality Metrics.” *RLG DigiNews* 4, no. 4, August (2000). www.rlg.org/preserv/diginews/diginews4-4.html#technical1, (acc. Oct 2000).

Williams, D. “Interpreting Digital Scanner/Camera Specifications.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 40-41.

Williams, D. “What is an MTF...and Why Should You Care?” *RLG DigiNews* 2, no. 1, February 15 (1998). www.rlg.org/preserv/diginews/diginews21.html#technical, (acc. Oct 2000).

Williams, J. “Guidelines for Image Capture: a UK Perspective.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 61-71.

X-Rite®. *The Color Guide and Glossary: Communication, Measurement, and Control for Digital Imaging and Graphic Arts.* 1998. www.x-rite.com/Documents/Mktg/L11-029.pdf, (acc. Oct 2000).

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

9. Asset Management

“"New Library: the People’s Network": the Government’s response, Cm 3887.” 1998, HMSO, www.culture.gov.uk/new-library.htm, (acc. Oct 2000).

“Archives On-Line: The Establishment of a United Kingdom Archival Network.” National Council on Archives, www.archives.org.uk/index2.html, (acc. Oct 2000), (created 1998).

“Archiving and Distribution Working Group Charter.” GII Virtual Roundtable, www.globalcollaboration.org/jtc1/igbt/adwg.html, (acc. Oct 2000), (created Nov 1999*).

“Canto Software home page.” www.canto-software.com, (acc. Jan 2000).

“COLLAGE-project by Guildhall Library.” Corporation of London, collage.nhil.com/index.htm, (acc. Oct 2000), (created 1997).

“Creation of the Images and Database.” Historic American Sheet Music, Rare Books, Manuscript and Special Collections Library, scriptorium.lib.duke.edu/sheetmusic/techinfo-scanning.html, (acc. Oct 2000), (created 1999).

“Data Standards.” Visual Resources Association, www.oberlin.edu/~art/vra/dsc.html, (acc. Oct 2000), (created June 2000).

“DC Date Qualifiers: DC (Final) Working Draft - 14 December 1999.” 1999, Dublin Core Metadata Initiative, www.mailbase.ac.uk/lists/dc-usage/files/date-final.html, (acc. Oct 2000), (created December 1999).

“Digital Libraries: Metadata Resources.” IFLANET, www.ifla.org/II/metadata.htm, (acc. Oct 2000), (created Oct 2000*).

“DOI: The Digital Object Identifier home page.” The Digital Object Identifier System, www.doi.org, (acc. Oct 2000), (created Sept 2000*).

“Dublin Core Metadata Initiative home page.” purl.oclc.org/dc, (acc. Oct 2000), (created Oct 1998).

“Ein neues Konvertierungsprogramm der SUB Göttingen/GDZ zur Umwandlung bibliographischer Pica/GBV-Daten in RDF/XML.” Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/gdz-tools/pica2xml/pica2xml_index_de.html, (acc. Oct 2000), (created 1997).

“EmbARK, Image Management Module. Version 1.2. Review.” *Visual Resources* XII (1996): 81-93.

“Guidance on expressing the Dublin Core within the Resource Description Framework (RDF).” Dublin Core Metadata Initiative, www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/, (acc. Oct 2000), (created July 1999*).

“Guidelines and Standards: Getting Started.” Colorado Digitization Project, coloradodigital.coalliance.org/standard.html, (acc. Jan 2000), (created 1999).

“Image Scanning.” Harvard University, Library Digital Initiative, hul.harvard.edu/ldi/html/image_scanning.html, (acc. Oct 2000), (created 1998).

“Indexing and Morphology: Why You Need Morphology.” Lingsoft Inc., www.lingsoft.fi/doc/indexing/morph.html, (acc. Oct 2000), (created Dec 1994).

“JSTOR Usage Statistics.” Journal Storage, stats.jstor.org, (acc. Oct 2000), (created 2000).

“KODAK PHOTO CD Technology/Transferring Images.” *Kodak Service & Support Frequently Asked Questions,* www.kodak.com/cluster/global/en/service/faqs/faq1001c.shtml, (acc. Oct 2000), (created Nov 1998*).

“Lessons Learned: National Digital Library Competition.” The Library of Congress, lcweb2.loc.gov/ammem/award/lessons/lessons.html, (acc. Oct 2000), (created April 1999).

“Library of Congress Maintenance Agency Page for International Standard Z39.50.” Library of Congress, lcweb.loc.gov/z3950/agency, (acc. Oct 2000), (created Aug 2000*).

“Making of America II.” *The Making of America Testbed Project White Paper. Version 2.0, February 15,* 1998, http://sunsite.berkeley.edu/moa2/wp-v2.html, (acc. Oct 2000), (created March 1998).

“MARC 21 Concise Bibliographic: Control Field 007.” Library of Congress, lcweb.loc.gov/marc/bibliographic/ecbd007s.html#mrcb007c, (acc. Oct 2000), (created Jan 2000*).

“MARC Standards.” Library of Congress, Network Development & MARC Standards Office, lcweb.loc.gov/marc, (acc. Oct 2000), (created June 2000*).

“Museums and the Online Archive of California.” Online Archive of California, www.oac.cdlib.org/dynaweb/ead/moac/, (acc. Oct 2000).

“NEDLIB - National Electronic Deposit Library - EU-funded project to develop a European library of digital documents in a stable archival environment.” www.kb.nl/coop/nedlib/, (acc. Aug 2000), (created Jan 1998).

“OCR and Text Recognition: Commercial Research and Products.” Video and Image Data Exploration and Organization (VIDEO), hera.itc.it:3003/~messelod/OCR/Products.html, (acc. Oct 2000).

“Optical Image Technology, Inc. home page.” www.opticaltech.com, (acc. Oct 2000).

“Perseus Project home page.” Tufts University, www.perseus.tufts.edu, (acc. Oct 2000).

“Phillips Document Management Source Book.” The Business Information Library, www.biz-lib.com/ZPHDMS.html, (acc. Oct 2000), (created Sept 2000*).

“Resource Description Framework (RDF).” Word Wide Web Consortium, www.w3.org/RDF/, (acc. Oct 2000).

“RLG Working Group on Preservation Issues of Metadata: Final Report.” Research Libraries Group, www.rlg.org/preserv/presmeta.html, (acc. Oct 2000), (created May 1998).

“Sample Structural Metadata for the Coolidge Image-Only Documents.” *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/timag-samp.html, (acc. Oct 2000), (created March 1997).

“Sample Structural Metadata for the Coolidge Photos.” *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/photo-samp.html, (acc. Oct 2000), (created March 1997).

“Sample Structural Metadata for the Coolidge SGML Encoded Documents.” *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/sgml-samp.html, (acc. Oct 2000), (created April 1997).

“The Conversion Process.” Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html#conversion, (acc. Oct 2000), (created June 1999).

“The Data Documentation Initiative: A Project to Develop an XML Document Type Definition for Data Documentation.” ICPR, www.icpsr.umich.edu/DDI/code-book.html, (acc. Oct 2000), (created March 2000).

“Towards a Learning Nation: the Digital Contribution Recommendations Proposed by the Federal Task Force on

Digitization, Final Report, 1997.” National Library of Canada, www.nlc-bnc.ca/coopprog/finalreport/eindex.htm, (acc. Oct 2000), (created Dec 1997).

“Visual Information Access system (VIA).” Harvard University Library, hul.harvard.edu/ldi/html/via_metadata.html, (acc. Oct 2000), (created Feb 1999).

Acland, G., K. Cumming, and S. McKemmish. “The End of the Beginning: The SPIRT Recordkeeping Metadata Project.” *Australian Society of Archivists 1999 Conference,* 1999, www.archivists.org.au/events/conf99/spirt.html, (acc. Oct 2000).

Alkhoven, P. “Metadata Needed for Handling Digital Content.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 35-41.

Arms, C., J. Klavans, and D. J. Waters. *Enabling Access in Digital Libraries: A Report on a Workshop on Access Management.* Washington, DC: The Digital Library Federation, 1999.

Arms, W. Y. “Implementing Policies for Access Management.” *D-Lib Magazine* 4, no. 2, February (1998). www.dlib.org/dlib/february98/arms/02arms.html, (acc. Oct 2000).

Aronoff, S. *Geographical Information Systems: A Management Perspective.* Ottawa: WDL Publications, 1989.

Atkins, H., C. Lyons, H. Ratner, C. Risher, C. Shillum, D. Sidman, and A. Stevens. “Reference linking with DOIs: a case study.” *D-Lib Magazine* 6, no. 2, February (2000). www.dlib.org/dlib/february00/02risher.html, (acc. Oct 2000).

Bainbridge, D., and T. C. Bell. “An extensible optical music recognition system.” In *Proceedings of the 1996 Australasian Computer Science Conference*, Melbourne, Australia, 1996. 308-317.

Band, J., and J. S. Gowdy. “S*ui Generis* Database Protection: Has Its Time Come?” *D-Lib Magazine* 3, no. 6, June (1997). www.dlib.org/dlib/june97/06band.html, (acc. Oct 2000).

Barber, D. “OhioLINK: A Consortial Approach to Digital Library Management.” *D-Lib Magazine* 3, no. 4, April (1997). www.dlib.org/dlib/april97/04barber.html, (acc. Oct 2000).

Batty, D. “WWW - Wealth, Weariness or Waste: Controlled vocabulary and thesauri in support of online information access.” *D-Lib Magazine* 4, no. 11, November (1998). www.dlib.org/dlib/november98/11batty.html, (acc. Oct 2000).

Beagrie, N., and D. Greenstein. “A Strategic Policy Framework for Creating and Preserving Digital Collections.” 1998, Arts and Humanities Data Service, www.ahds.ac.uk/manage/framework.htm, (acc. Jan 2000).

Beagrie, N., and D. Greenstein. “Managing Digital Collections: AHDS Policies, Standards and Practices Version 1.” 1998, Arts and Humanities Data Service, www.ahds.ac.uk/public/srg.html, (acc. Oct 2000), (created Dec 1998).

Bearman, D. “New Economic Models for Administering Cultural Intellectual Property.” 1996, Archives and Museums Informatics, www.archimuse.com/papers/db.mesl/economics.html, (acc. Oct 2000), (created Feb 1996).

Bearman, D. “NISO/CLIR/RLG Technical Metadata Elements for Images Workshop, April 18-19, 1999.” NISO (National Information Standards Organization), www.niso.org/imagerpt.html, (acc. Oct 2000), (created April 1999).

Bearman, D. “Optical media: their implications for archives and museums.” In *Archives and Museum Informatics Technical Report, 1(1)*, Pittsburgh, PA: Archives and Museum Informatics, 1987.

Bearman, D., and J. Trant. “Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process.” *D-Lib Magazine* 4, no. 6, June (1998). www.dlib.org/dlib/june98/06bearman.html, (acc. Oct 2000).

Bearman, D., and K. Sochats. “Metadata Requirements for Evidence.” University of Pittsburgh, www.lis.pitt/edu/~nhprc/BACartic.html, (acc. Oct 2000).

Bearman, D., and M. Hedstrom. “Reinventing archives for electronic records: alternative service delivery

options." In *Electronic records management program strategies*, edited by M Hedstrom, Pittsburgh, PA: Archives and Museum Informatics, 1993. 82-98.

Bearman, D., G. Rust, S. Weibel, E. Miller, and J. Trant. "A Common Model to Support Interoperable Metadata. Progress Report on Reconciling Metadata Requirements from the Dublin Core and INDECS/DOI Communities." *D-Lib Magazine* 5, no. 1, January (1999). www.dlib.org/dlib/january99/bearman/01bearman.html, (acc. Oct 2000).

Beavers, A. F. "Evaluating Search Engine Models for Scholarly Purposes: A Report from the Internet Applications Laboratory." *D-Lib Magazine* 4, no. 12, December (1998). www.dlib.org/dlib/december98/12beavers.html, (acc. Oct 2000).

Benemann, W. E. "Reference Implications of Digital Technology in a Library Photograph Collection." *Reference Services Review* 22, no. 4, Winter (1994): 45-50.

Benton Foundation. "Buildings, Books, and Bytes: Libraries and Communities in the Digital Age." www.benton.org/Library/Kellogg/buildings.html, (acc. Oct 2000), (created Nov 1996).

Berger, B. "Digital Imaging Projects at Cornell University." In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 49-57.

Berghel, H., and L. O'Gorman. "Protecting ownership rights through digital watermaking." *Computer* 7, July 29 (1996): 101-103.

Bergman, L. D., V. Castelli, and C.-S. Li. "Progressive Content-Based Retrieval from Satellite Image Archives." *D-Lib Magazine* 3, no. 10, October (1997). www.dlib.org/dlib/october97/ibm/10li.html, (acc. Oct 2000).

Bescós Ramón, J. "Image Processing Algorithms for Readability Enhancement of Old Manuscripts." *Electronic Imaging* 89, no. 1 (1989): 392-97.

Bescós, J., and J. Navarro. "La Digitalización como Medio para la Preservación y Acceso a la Información en Archivos y Bibliotecas." *Educación y Bibliotecas* 80 (1987): 28-41.

Besser, H. "Adding an Image Database to an Existing Library and Computer Environment: Design and Technical Considerations." In *ASIS 1991 Midyear Proceedings: Multimedia Information Systems 1992*, 1992.

Besser, H. "User Interfaces for Museums." *Visual Resources* 7 (1991): 293-309.

Besser, H. "Visual Access to Visual Images: The UC Berkeley Image Database Project." *Library Trends* 38, no. 4 (1990): 787-798.

Besser, H., and J. Trant. *Introduction to Imaging: Issues in Constructing an Image Database.* Santa Monica, CA: The Getty Art History Information Program, 1995. www.getty.edu/gri/standard/introimages/index.html, (acc. Oct 2000).

Besser, H., and M. Snow. "Access to Diverse Collection in a University Setting: The Berkeley Dilemma." In *Beyond the Book: Extending MARC for Subject Access to Non-Traditional Materials*, edited by Pat Moholt and Toni Petersen, New York: G. K. Hall & Co, 1990.

Bewley, R., D. Donaghue, V. Gaffney, M. van Leusen, and A. Wise, (eds.) *Archiving Aerial Photography and Remote Sensing Data: A Guide to Good Practice*. 1999, Oxford: Oxbow Books for the Arts and Humanities Data Service.

Bicknese, D. "Measuring the Accuracy of the OCR in the Making of America." *(a report prepared in fulfillment of directed field experience requirements, University of Michigan, School of Information, winter 1998),* University of Michigan, moa.umdl.umich.edu/moaocr.html, (acc. Oct 2000), (created 1998).

Bikson, T. K. "Managing Digital Documents: Technology Challenges and Institutional Responses." In *Proceedings of the International Conference of the Round Table on Archives*, Stockholm, Sweden: International Council on Archives, *forthcoming*.

Birdsey, C. L. "Creating Accessible Digital Imagery." *D-Lib Magazine* 6, no. 4, April (2000). www.dlib.org/dlib/april00/birdsey/04birdsey.html, (acc. Oct 2000).

Borgman, C. L. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World.* Cambridge, MA: MIT Press, 2000.

Braudaway, G., K. Magerlein, and F. Mintzer. “Protecting publicly available images with a visible image watermark.” In *Proceedings of Optical Security and Counterfeit Deterrence Techniques*, Bellingham, WA: SPIE, 1996. 287-289.

Brett, G. H., II. “An Integrated System for Distributed Information Services.” *D-Lib Magazine* 2, no. 12, December (1996). www.dlib.org/dlib/december96/dipps/12brett.html, (acc. Oct 2000).

Breu, M., A. Brüggermann-Klein, C. Haber, and R. Weber. “The MeDoc distributed electronic library - Accounting and security aspects.” In *Proceedings of the ICCC/IFIP Conference (April 1997, Canterbury, England)*, Canterbury: University of Kent, 1997.

Buckland, M., A. Chen, H.-M. Chen, Y. Kim, B. Lam, R. Larson, B. Norgard, and J. Purat. “Mapping Entry Vocabulary to Unfamiliar Metadata.” *D-Lib Magazine* 5, no. 1, January (1999). www.dlib.org/dlib/january99/buckland/01buckland.html, (acc. Oct 2000).

Cawkwell, A. E. “Developments in Indexing Picture Collection.” *Information Services and Use* 13, no. 4 (1993): 381-388.

Chen, H. “Semantic Research for Digital Libraries.” *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/chen/10chen.html, (acc. Oct 2000).

Chilvers, A., and J. Feather. “The Management of Digital Data: A Metadata Approach.” *The Electronic Library* 16, no. 6, December (1998): 367-372.

Cisco, S. L., and J. Wertzberger. “Indexing Digital Documents: It’s NOT an Option - Pay now or Pay (More) Later!” *Inform* February (1997).

Clews, J. “Digital Language Access: Scripts, Transliteration, and Computer Access.” *D-Lib Magazine* 3, no. 3, March (1997). www.dlib.org/dlib/march97/sesame/03clews.html, (acc. Oct 2000).

Cohen, E. “Internet2 and the Next Generations.” *NARAS Journal* 8, no. 2, Winter (1998/99): 51-56.

Cohen, J. E. “Some reflections on copyright management systems and laws designed to protect them.” *Berkeley Tech. Law Journal* 12, no. 1 (1997): 161-188.

Collections, Content, and the Web. Washington, DC: Council on Library and Information Resources, 2000. www.clir.org/cpa/reports/pub88/contents.html, (acc. Oct 2000).

Cox, J., and M. Taleb. “Images on the Internet: Enhanced User Access.” *Database* 17, no. 4, August (1994): 18-22, 24-26.

Coyle, K. “The Virtual Union Catalog: A Comparative Study.” *D-Lib Magazine* 6, no. 3, March (2000). www.dlib.org/dlib/march00/coyle/03coyle.html, (acc. Oct 2000).

Cranshaw, T. “Delivering Images.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 12-16.

Crist, M., and J. Price-Wilkin. “The Access Context.” In *Selecting Library and Archive Collections for Digital Reformatting: Proceedings from an RLG Symposium held November 5-6, 1995, in Washington, DC*, edited by Ricky Erway, Mountain View, CA: RLG, 1996. 29-38.

Croft, W. B. “What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems).” *D-Lib Magazine* 1, no. 11, November (1995). www.dlib.org/dlib/november95/11croft.html, (acc. Oct 2000).

Dale, R. “Lossy or Lossless? File Compression Strategies Discussion at ALA.” *RLG DigiNews* 3, no. 2, February 15 (1999). www.rlg.org/preserv/diginews/diginews3-1.html#technical1, (acc. Oct 2000).

Daniel, R., and C. Lagoze. “Extending the Warwick Framework: From Metadata Containers to Active Digital Objects.” *D-Lib Magazine* 3, no. 11, November (1997). www.dlib.org/dlib/november97/daniel/11daniel.html, (acc. Oct 2000).

Dartois, M., A. Maeda, T. Sakaguchi, T. Fujita, S. Sugimoto, and K. Tabata. “A multi-lingual electronic text collection of folk tales for casual users using off-the-shelf browsers.” *D-Lib Magazine* 3, no. 10, October (1997). www.dlib.org/dlib/ocotber97/sugimoto/10sugimoto.html, (acc. Oct 2000).

Dartois, M., T. Sakaguchi, S. Sugimoto, K. Tabata, A. Maeda, and T. Fujita. “Building a multi-lingual electronic text collection of folk tales as a set of encapsulated document objects: An approach for casual users to browse multi-lingual documents on the fly.” In *Research and Advanced Technology for Digital Libraries: First European Conference, ECDL ’97, Pisa, Italy, September 1-3, 1997*, edited by Carol Peters and Costantino Thanos, Springer, New York, 1997. 215-231.

Das, M., R. Manmatha, and E. M. Riseman. “Indexing flower patent images using domain knowledge.” *IEEE Intelligent Systems* 14, no. 5 (September/October) (1999): 24-33.

Day, M. “Extending Metadata for Digital Preservation.” *Ariadne* 9, no. May (1997). www.ariadne.ac.uk/issue9/metadata/intro.html, (acc. Oct 2000).

Dementi, M. R. “Access and Archiving as a New Paradigm.” *The Journal of Electronic Publishing* 3, no. 3, March (1998). www.press.umich.edu/jep/03-03/dementi.html, (acc. Oct 2000).

Dempsey, L., and S. L. Weibel. “The Warwick Metadata Workshop: a framework for the deployment of resource description.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/07weibel.html, (acc. Oct 2000).

Digitisation and the challenge to copyright: CLIP seminar Thursday 28 April, 1994, *CLIP seminar report*. 1994: Intellectual Property Institute.

Dovey, M. “Meta-Objects: An Object Orientated Approach to Metadata.” *Ariadne* 19. www.ariadne.ac.uk/issue19/meta-objects/, (acc. Oct 2000).

Eakins, J. P. “Content-Based Image Retrieval.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 114-115.

Eakins, J. P., and M. E. Graham. “Content-Based Image Retrieval: A Report to the JISC Technology Applications Programme.” 1999, www.unn.ac.uk/iidr/report.html, (acc. Oct 2000), (created Jan 1999).

Edelson, D. C., and D. N. Gordin. “Adapting Digital Libraries for Learners: Accessibility vs. Availability.” *D-Lib Magazine* 2, no. 9, September (1996). www.dlib.org/dlib/september96/nwu/09edelson.html, (acc. Oct 2000).

Elachi, C. “Radar Images from Space.” *Scientific American* 247, no. 6 (1982): 54-61.

Enser, P. G. B. “Progress in Documentation: Pictorial Information Retrieval.” *Journal of Documentation* 51, no. 2 (1995): 126-170.

Ester, M. “Building Towards Image Resources and User Environments in the Visual Arts.” In *Proceedings of the 1993 Electronic Imaging and the Visual Arts Conference: EVA ’93*, edited by James Hemsley, Neil Sandford and David Saunders, Aldershots, Hants: Braumer Ltd., 1993.

Ester, M. “Issues in the Use of Electronic Images for Scholarship in the Arts and Humanities.” In *Networking in the Humanities (proceedings of the Second Conference on Scholarship and Technology in the Humanities, held at Elvetham Hall, Hampshire, UK, 13-16 April, 1994: papers in honour of Michael Smethurst for his 60th birthday)*, edited by S. Kenna and S. Ross, London: Bowker-Saur, 1995. 111-125.

Eyre, J. “ELISE II: Networked Delivery of Images and User Interface Issues.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 31-34.

Feder, J. “Towards Image Content-Based Retrieval for the World-Wide Web.” *Advanced Imaging* 11, no. 1 (1996): 26-29.

Federal Task Force on Digitization (Canada). *Towards a learning nation: The digital contribution: recommendations proposed by the Federal Task Force on Digitization: final report.* Ottawa: National Library of Canada, 1999.

Ferguson, I. A., and M. J. Wooldridge. “Paying Their Way: Commercial Digital Libraries for the 21st Century.” *D-*

Lib Magazine 3, no. 6, June (1997). www.dlib.org/dlib/june97/zuno/06ferguson.html, (acc. Oct 2000).

Fisher, A. “Content provision in the digital era.” *Assignation: ASLIB Social Sciences Information Group Journal* 15, no. 1 (1997): 4-9.

Flickner, M., H. Swahney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. “Query by Image and Video Content: The QBIC System.” *IEEE Computer* 28, no. 9 (1995): 23-32.

Flickschuh, H. “Laurin: Technical Presentation/Improx System.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 26-27.

Forsyth, D. A., J. Malik, T. K. Leung, C. Bregler, C. Carson, H. Greenspan, and M. M. Fleck. “Finding Pictures of Objects in Large Collections of Images.” In *Digital Image Access and Retrieval: 1996 Clinic on Library Applications of Data Processing*, edited by P B Heidorn and Beth Sandore, Urbana, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1997. 118-139.

Fresko, M., and L. Brindley. *Optical discs and European libraries: a study of user and technical requirements.* London: Bowker Saur, 1992.

Frew, J., M. Freeston, R. B. Kemp, J. Simpson, T. Smith, A. Wells, and Q. Zheng. “The Alexandria Digital Library Testbed.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/alexandria/07frew.html, (acc. Oct 2000).

Frey, F., and S. Süsstrunk. “Color Issues to Consider in Pictorial Image Databases.” In *Proceedings, IS&T Fifth Color Imaging Conference*, Scottsdale, AZ, 1997. 112-115.

Gagos, T., with, P. Daub, A. Loftus, and S. Zachary. “Scanning the Past: A Modern Approach to Ancient Culture.” *Library Hi Tech* 14, no. 1 (1996): 11-22.

Galassi, P. “The Modern('s) Digital Archive.” *Photo District News*, no. October (1998).

Gay, G., R. Rieger, and A. Sturgill. “Findings of the MESL Casual User Survey.” In *Delivering Digital Images: Cultural Heritage Resources for Education, The Museum Educational Site Licensing Project*, edited by Christie Stephenson and Patricia McClung, Los Angeles: The Getty Information Institute, 1998.

Gillings, M., and A. Wise, (eds.) *GIS Guide to Good Practice*. 1999, Oxford: Oxbow Books for the Arts and Humanities Data Service.

Gilman, P. B., E. McInerney, and R. Vogel. “Some Simple Techniques for Improving the Print Quality of Digital Images from Kodak Electronic Cameras.” In *International Conference on Digital Printing Technologies, NIP 12*, Springfield, VA: IS&T, 1996. 227-232.

Gladney, H. M. “Safeguarding Digital Library Contents and Users: A Note on Universal Unique Identifiers.” *D-Lib Magazine* 4, no. 4, April (1998). www.dlib.org/dlib/april98/04gladney.html, (acc. Oct 2000).

Gladney, H. M. “Safeguarding Digital Library Contents and Users: Document Access Control.” *D-Lib Magazine* 3, no. 6, June (1997). www.dlib.org/dlib/june97/ibm/06gladney.html, (acc. Oct 2000).

Gladney, H. M. “Safeguarding Digital Library Contents and Users: Interim Retrospect and Prospects.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/gladney/07gladney.html, (acc. Oct 2000).

Gladney, H. M., and J. B. Lotspiech. “Safeguarding Digital Library Contents and Users: Assuring Convenient Security and Data Quality.” *D-Lib Magazine* 3, no. 5, May (1997). www.dlib.org/dlib/may97/ibm/05gladney.html, (acc. Oct 2000).

Gladney, H. M., and J. B. Lotspiech. “Safeguarding Digital Library Contents and Users: Storing, Sending, Showing, and Honoring Usage Terms and Conditions.” *D-Lib Magazine* 4, no. 5, May (1998). www.dlib.org/dlib/may98/gladney/05gladney.html, (acc. Oct 2000).

Gladney, H. M., F. Mintzer, and F. Schiattarella. “Safeguarding Digital Library Contents and Users. Digital Images of Treasured Antiquities.” *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/vatican/07gladney.html, (acc. Oct 2000).

Gladney, H. M., F. Mintzer, F. Schiattarella, J. Bescós, and M. Treu. “Digital Access to Antiquities.” *Communications of the ACM* 41, no. 4, April (1998): 49-57.

Gonçalves, M. A. “Use of Hypermedia Models in Digital Libraries for Geographic Data.” Master’s, UNICAMP, 1997.

González García, P. “New Technology and the Reconstruction of the Archival Heritage.” In *Proceedings of the XXX International Conference of the Round Table on Archives, Thessaloniki, 1994*, Dordrecht, The Netherlands: International Council on Archives, 1998. 125-29.

González, P. “Databases and Long Distance Communication. A Spanish Experience.” In *Proceedings of the Third European Conference on Archives. Vienna, 1993*: Sonderband, 1993. 319-72.

Gordon, C. “Pattern of User Queries in an ICONCLASS Database.” *Visual Resources* XII (1996): 177-186.

Gordon, M. E. “Image Databases in North America.” *Special issue of Computers in the History of Art* 4, no. 2 (1994).

Green, D. “Beyond Word and Image. Networking Moving Images: More Than Just the "Movies"." *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/07green.html, (acc. Oct 2000).

Greenstein, D., and P. Miller. “Discovering On-Line Resources Across the Humanities: A Practical Implementation of the Dublin Core.” 1997, ahds.ac.uk/public/metadata/discovery.html, (acc. Oct 2000).

Grout, C. “COLDIP: The Corporation of London Data Imaging Project. One of a New Age of Imaging Projects.” Visual Arts Data Service, http://vads.ahds.ac.uk/training_advice/case_studies/Coldip.html, (acc. Oct 2000), (created Nov 1997).

Gudivada, V. N., and V. V. Raghavan. “Content Based Image Retrieval Systems.” *IEEE Computer* 28, no. 9 (1995): 18-22.

Guidelines for Digital Imaging: Papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998. London: The National Preservation Office, 1998. www.rlg.ac.uk/preserv/joint/, (acc. Oct 2000).

Hakala, J. “Metadata and identifiers for e-journals.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 53-68.

Handel, M. “Issues of Scale and Size in Visual Databases.” 1995, University of California at Berkeley, sunsite.berkeley.edu/Imaging/Databases/Fall95papers/handel.html, (acc. Oct 2000), (created Sept 1996*).

Handel, M. “Proposed Interface for Scale Information.” 1995, University of California at Berkeley, http://sunsite.berkeley.edu/Imaging/Databases/Fall95papers/handel2.html, (acc. Oct 2000), (created Sept 1996*).

Hert, C. A., and G. Marchionini. “Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations.” 1987, University of North Carolina, http://ils.unc.edu/~march/blsreport/mainbls.html, (acc. Oct 2000), (created July 1997).

Herzberg, A. “Safeguarding Digital Library Contents: Charging for Online Content.” *D-Lib Magazine* 4, no. 1, January (1998). www.dlib.org/dlib/january98/ibm/01herzberg.html, (acc. Oct 2000).

Hillmann, D. I. “Choices: MARC or Dublin Core?” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 89-90.

Hirtle, P. B. “Image Management Systems and Web Delivery.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 119-134.

Hofman, P., E. Worsfold, D. Hiom, M. Day, and A. Oehler. “Specification for Resource Description Methods: 2, Selection Criteria for Quality Controlled Information Gateways.” *DESIRE: Development of a European Service for Information on Research and Education, Deliverable 3.2(2),* 1997, www.ukoln.ac.uk/metadata/desire/quality, (acc. Oct 2000), (created May 1997).

Holt, B., and L. Hartwick. “Retrieving Art Images by Image Content: The UC Davis QBIC Project.” In *Electronic Library and Visual Information Research, ELVIRA 1: Proceedings of the First ELVIRA Conference, Held in May 1994 at De Montfort University, Milton Keynes*, edited by Mel Collier and Kathryn Arnold, London: Aslib, 1995. 93-100.

Hopkin, D. “Shifting the Focus: Digital Imaging & Photographic Collections Management at the National Railway Museum.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 72-80.

Jenkins, F. R., T. A. Nartker, and S. V. Rice. “Testing OCR Accuracy: Results of the Fifth Annual Test of OCR Technology by UNLV’s Information Science Research Institute.” *INFORM* September 10, no. 8 (1996): 20-25.

Johnson, N. F., and S. Jajodia. “Exploring Steganography: Seeing the Unseen.” *Computer* February 31, no. 2 (1998): 26-34.

Jones, M. L. W., G. K. Gay, and R. H. Rieger. “Project Soup: Comparing Evaluations of Digital Collection Efforts.” *D-Lib Magazine* 5, no. 11, November (1999). www.dlib.org/dlib/november99/11jones.html, (acc. Oct 2000).

Kahn, R., and R. Wilensky. “A Framework for Distributed Object Services.” 1995, The Corporation for National Research Initiatives (CNRI), www.cnri.reston.va.us/home/cstr/arch/k-w.html, (acc. Oct 2000), (created May 1995).

Kaser, R. T. “If Information Wants to Be Free . . . Then Who’s Going to Pay for It?” *D-Lib Magazine* 6, no. 5, May (2000). www.dlib.org/dlib/may00/kaser/05kaser.html, (acc. .

Ketelaar, E. “Exploitation of New Archival Materials.” In *Proceedings of the 11th International Congress on Archives*, Munich: K. G. Sauer, 1989. 189-99.

Kiernan, K. “Digital Preservation, restoration, and the dissemination of medieval manuscripts.” In *Gateways, gatekeepers, and roles in the information omniverse: proceedings of the third symposium: November 13-15, 1993*, edited by A Okerson and D Mogge, Washington, DC, 1994.

Kohl, U., J. Lotspiech, and M. A. Kaplan. “Safeguarding Digital Library Contents and Users: Protecting Documents Rather Than Channels.” *D-Lib Magazine* 3, no. 9, September (1997). www.dlib.org/dlib/september97/ibm/09lotspiech.html, (acc. Oct 2000).

Lagoze, C. “A Secure Repository Design for Digital Libraries.” *D-Lib Magazine* 1, no. 12, December (1995). www.dlib.org/dlib/december95/12lagoze.html, (acc. Oct 2000).

Lagoze, C. “From Static to Dynamic Surrogates: Resource Discovery in the Digital Age.” *D-Lib Magazine* 3, no. 6, June (1997). www.dlib.org/dlib/june97/06lagoze.html, (acc. Oct 2000).

Lagoze, C. “Metadata Creation: From Digital Images to a Database of Documents.” In *Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Stephen Chapman, Ithaca, NY: Cornell University Library, 1996.

Lagoze, C., and D. J. Dienst. “An architecture for distributed document libraries.” *Communications of the ACM* 38, no. 4, April (1995): 47.

Lagoze, C., and S. Payette. “Metadata: Principles, Practices, and Challenges.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 84-100.

Lancaster, F. W. *Vocabulary Control for Information Retrieval.* Arlington, VA: Information Resources Press, 1986.

Lougee, W. P. “PEAK: Pricing Electronic Access to Knowledge.” *(paper presented at Coalition for Networked Information),* 1999, University of Michigan, www.lib.umich.edu/libhome/DLI/presentations/cni99/sld001.htm, (acc. Oct 2000), (created April 1999).

Lunin, L. F. “The Big Picture: Selection and Design Issues for Image Information Systems.” In *Digital Image Access & Retrieval*, edited by P Bryan Heidorn and Beth Sandore, Urbana, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1997.

Lyons, P. A. “Access to Digital Objects: A Communications Law Strategy.” *D-Lib Magazine* 1, no. 10, October (1995). www.dlib.org/dlib/october95/10lyons.html, (acc. Oct 2000).

Lytle, R. H. “Intellectual Access to Archives.” *American Archivist* 43 (1980): 64-76.

MacIntyre, R., and S. Tanner. “*Nature* - A Prototype Digital Archive.” *(paper presented at the sixth DELOS Workshop Preservation of Digital Information, June 17-19, 1998),* 1998, Higher Education Digitisation Service, heds.herts.ac.uk/resources/papers/HEDSnature.pdf, (acc. Oct 2000), (created Jun 1998).

MacKie-Mason, J. K., M. S. Bonn, J. F. Riveros, and W. P. Lougee. “A Report on the PEAK Experiment: Usage and Economic Behaviour.” *D-Lib Magazine* 5, no. 7/8 (1999). www.dlib.org/dlib/july99/mackie-mason/07mackie-mason.html, (acc. Oct 2000).

Maly, K., M. L. Nelson, and M. Zubair. “Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework.” *D-Lib Magazine* 5, no. 3, March (1999). www.dlib.org/dlib/march99/maly/03maly.html, (acc. Oct 2000).

MARC, The Machine Readable Catalogue Record, Specified as ISO 2709.

Marchionini, G., C. Plaisant, and A. Komlodi. “Interfaces and Tools for the Library of Congress National Digital Library Program.” *Information Processing and Management* 34, no. 5 (1998): 535-555.

Matthews, L. “The Collection Management Perspective.” In *Selecting Library and Archive Collections for Digital Reformatting: Proceedings from an RLG Symposium held November 5-6, 1995, in Washington, DC*, edited by Ricky Erway, Mountain View, CA: RLG, 1996. 100-101.

McGrath, R. E. “Caching for Large Scale Systems: Lessons from the WWW.” *D-Lib Magazine* January (1996). www.dlib.org/dlib/january96/ncsa/01mcgrath.html, (acc. Oct 2000).

Miller, E. “An Introduction to the Resource Description Framework.” *D-Lib Magazine*, no. May (1998). www.dlib.org/dlib/may98/miller/05miller.html, (acc. Oct 2000).

Miller, P. “Metadata for the Masses.” *Ariadne* 5. www.ariadne.ac.uk/issue5/metadata-masses/intro.html, (acc. Oct 2000).

Millman, D. “Cross-Organizational Access Management: A Digital Library Authentication and Authorization Architecture.” *D-Lib Magazine* 5, no. 11, November (1999). www.dlib.org/dlib/november99/millman/11millman.html, (acc. Oct 2000).

Mintzer, F. C., L. E. Boyle, A. N. Cazes, B. S. Christian, S. C. Cox, F. P. Giordano, H. M. Gladney, J. C. Lee, M. L. Kelmanson, A. C. Lirani, K. A. Magerlein, A. M. B. Pavani, and F. Schiattarella. “Toward On-line, Worldwide Access to Vatican Library Materials.” *IBM Journal of Research and Development* 40, no. 2 (1996). www.research.ibm.com/journal/rd/mintz/mintzer.html, (acc. Oct 2000).

Mintzer, F., J. Lotspiech, and N. Morimoto. “Safeguarding Digital Library Contents and Users: Digital Watermarking.” *D-Lib Magazine* 3, no. 12, December (1997). www.dlib.org/dlib/december97/ibm/12lotspiech.html, (acc. Oct 2000).

Moen, W. E. “Accessing Distributed Cultural Heritage Information.” *Communications of the ACM* 41, no. 4, April (1998): 45-48.

Museum Documentation Association. *MDA Archaeological Objects Thesaurus.* Cambridge: Museum Documentation Association, English Heritage, and the Royal Commission for the Historical Monuments of England, 1997.

Museum Documentation Association. *SPECTRUM: The UK Museum Documentation Standard.* 2nd ed. Cambridge: Museum Documentation Association, 1997.

Myaeng, S. “MIRAGE: A prototype for multimedia information retrieval and gathering environment.” In *Proceedings of the International Conference on Digital Libraries and Information Services for the 21st Century (Seoul, Sept 10-13)*, Seoul: Hankook Computer Industrial Co., 1996. 115-125.

Myaeng, S., D. Shin, W. Chun, and K. Lee. “MIRAGE: A prototype for a multimedia digital library.” *Journal of the Korean Information Science Society* 3, no. 6, December (1997): 605-618.

National Archives and Records Administration (NARA). *Digital Imaging and Optical Digital Data Disk Storage Systems: Long-Term Access Strategies for Federal Agencies.*, *Technical Information Paper no. 12*: NARA, 1994.

National Council on Archives. *Archives On-Line.* London: National Council on Archives, 1999.

National Council on Archives. *British Archives: The Way Forward.* London: National Council on Archives, 1998.

National Council on Archives. *Rules for the construction of personal, place and corporate names.*: NCA, 1997.

Nevill-Manning, C. G., I. H. Witten, and G. W. Paynter. "Browsing in Digital Libraries: A phrase-based approach." In *Proceedings of ACM Digital Libraries*, 1997. 230-236.

Niblack, W., and M. Flickner. "Find me the Pictures that Look Like This: IBM's Image Query Project." *Advanced Imaging* April (1993): 32-35.

Noerr, P. "The Digital Library Toolkit." *Second edition,* 2000, Sun Microsystems, www.sun.com/products-n-solutions/edu/libraries/digitaltoolkit.html, (acc. Oct 2000), (created March 2000).

Oard, D. W. "Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries." *D-Lib Magazine* 3, no. 12, December (1997). www.dlib.org/dlib/december97/oard/12oard.html, (acc. Oct 2000).

Oard, D., C. Peters, M. Ruiz, R. Frederking, J. Klavans, and P. Sheridan. "Multilingual Information Discovery and AccesS (MIDAS): A Joint ACM DL'99 / ACM SIGIR'99 Workshop." *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/10oard.html, (acc. Oct 2000).

Okon, C. "IBMs Image Recognition Technology for Databases at Work: QBIC or Not QBIC?" *Advanced Imaging* May (10) (1995): 63-65.

Paskin, N. "DOI: Current Status and Outlook." *D-Lib Magazine* 5, no. 5, May (1999). www.dlib.org/dlib/may99/05paskin.html, (acc. Oct 2000).

Paskin, N., and G. Rust. "The Digital Object Identifier Initiative: Metadata Implications." *DOI Discussion Paper 2, version 3,* 1999, dx.doi.org/10.100/104, (acc. Oct 2000), (created Feb 1999).

Payette, S. "Persistent Identifiers in the Digital Terrain." *RLG DigiNews* 2, no. 2, April 15 (1998). www.rlg.org/preserv/diginews/diginews22.html#Identifiers, (acc. Oct 2000).

Payette, S., and C. Lagoze. "Value-Added Surrogates for Distributed Content: Establishing a Virtual Control Zone." *D-Lib Magazine* 6, no. 6, June (2000). www.dlib.org/dlib/june00/payette/06payette.html, (acc. Oct 2000).

Pejterson, A. M. "Semantic Information Retrieval." *Communications of the ACM* 41, no. 4, April (1998): 90-92.

Perkins, J. "The CIMI standard for the interchange of multimedia information." In *Museums and Interactive Multimedia: Proceedings of an International Conference held in Cambridge, England, 20-24 September 1993 (6th International Conference on Hypermedia and Interactivity in Museums)*, edited by Diane Lees and Museum Documentation Association, Cambridge: MDA, 1993. 244-254.

Perlman, E., and I. Kallen. "Common Internet File Formats." 1995, www.matisse.net/files/formats.html, (acc. Oct 2000), (created Dec 1995*).

Phelps, T., and R. Wilensky. "Toward active, extensible, networked documents: Multivalent architecture and applications." In *Proceedings of DL'96*, New York: ACM Press, 1996. 100-108.

Popkin, J., and A. Cushman. *Integrated Document Management - Controlling a Rising Standard.* Stamford, Conn: Gartner Group, 1993.

Prescott, A., and M. Pratt. "Excalibur: Image-Based Text Storage and Searching." In *Towards the Digital Library: The British Library's 'Initiatives for Access' Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott, London: The British Library, 1998. 178-189.

Price, L., and A. Smith. *Managing Cultural Assets from a Business Perspective.* Washington, DC: Council on Library and Information Resources, 2000.

Price-Wilkin, J. "Just-in-Time Conversion, Just-in-Case Collections: Effectively Leveraging Rich Document Formats for the WWW." *D-Lib Magazine* 3, no. 5, May (1997).

www.dlib.org/dlib/may97/michigan/05pricewilkin.html, (acc. Oct 2000).

Rabbani, M. “Image Compression Fundamentals.” In *The Compression Experience: Proceedings of The Rochester Section of the SMPTE (Society of Motion Picture & Television Engineers) Tutorial*, Rochester, NY, 1995. 7-24.

Reed-Scott, J., (ed.) *Cataloging of resources digitized for preservation: a SPEC Kit*. 1999, Washington, DC: Association of Research Libraries, Office of Leadership and Management Services.

Roberts, H. E. ""'Second Hand Images: The Roles of Surrogates in Artistic and Cultural Exchange". Paper delivered as the Keynote Address for the Visual Resources Satellite Session of the XXVIIIth International Congress of the History of Art (CIHA), Berlin, July 1992.” *Visual Resources* 9, no. 4 (1993): 335-346.

Rui, Y., T. S. Huang, and S. Mehrotra. “Relevance Feedback Techniques in Interactive Content-Based Image Retrieval.” In *Storage and Retrieval for Image and Databases VI, SPIE Proceedings 3312*, edited by K Sethi and Ramesh C Jain, Bellingham, WA: SPIE, 1997. 25-36.

Rusch-Feja, D., and U. Siebeky. “Evaluation of Usage and Acceptance of Electronic Journals: Results of an Electronic Survey of Max Planck Society Researchers including Usage Statistics from Elsevier, Springer and Academic Press.” *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/rusch-feja/10rusch-feja-summary.html, (acc. Oct 2000).

Rust, G. “Metadata: The Right Approach.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/rust/07rust.html, (acc. Oct 2000).

Rust, G., and M. Bide. “The <indecs> Metadata Model.” *version no. 3,* 1999, indecs, www.indecs.org/results/archive.htm, (acc. Oct 2000), (created July 1999).

Sairamesh, J., C. Nikolaou, D. F. Ferguson, and Y. Yemini. “Economic Framework for Pricing and Charging in Digital Libraries.” *D-Lib Magazine* 2, no. 2, February (1996). www.dlib.org/dlib/february96/forth/02sairamesh.html, (acc. Oct 2000).

Sandore, B. “What Users Want from Digital Image Collections.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 4-5.

Sandore, B., and N. Shaik. “Findings of the Instructor/Student Survey.” In *Delivering Digital Images: Cultural Heritage Resources for Education, The Museum Educational Site Licensing Project*, edited by Christie Stephenson and Patricia McClung, Los Angeles: The Getty Information Institute, 1998.

Schantz, H. F. “Optical Character Recognition - The Mature Technology with the Brilliant Future.” *INFORM* July/August 10, no. 7 (1996).

Seal, A. “The Victoria and Albert Museum and Electronic Imagebanks: Image Conversion and User Reactions.” *Program* 29, no. 4 (1995): 379-395.

Skaggs, D. “Electronic Document Management Systems as Archival Image Repositories.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 131-132.

Stephenson, C., and P. McClung, (eds.) *Delivering Digital Images, Cultural Heritage Resources for Education*. Vol. 1, *The Museum Educational Site Licensing Project*. 1998, Los Angeles, California: The Getty Information Institute.

Stephenson, C., and P. McClung, (eds.) *Images Online: Perspectives on the Museum Educational Site Licensing Project*. Vol. 2, *The Museum Educational Site Licensing Project*. 1998, Los Angeles: The Getty Information Institute.

Swain, D. J. “Enhancing and reconstructing digitized handwriting.” MSc, Virginia Polytechnic Institute and State University, 2000.

Talja, S., R. Heinisuo, E.-L. Kasesniemi, H. Kemppainen, S. Luukkainen, K. Pispa, and K. Järvelin. “Discourse Analysis of User Requests.” *Communications of the ACM* 41, no. 4, April (1998): 93-94.

TASI (Technical Advisory Service for Images). “Building Image Archives.” TASI,

www.tasi.ac.uk/framework/capture/overview.html, (acc. Oct 2000).

Tochtermann, K., and T. Alders. “DogitaLS1: A Digital Library System Based on Hyper-G.” *D-Lib Magazine* 2, no. 10, October (1996). www.dlib.org/dlib/october96/texas/10tochtermann.html, (acc. .

Torre, D. S. “KSR: Keywording for Subject Retrieval.” *Art Documentation* 14, no. 2 (1995): 29-36.

Valauskas, E. J. “Digital Images over the Internet: Rome Reborn at the Library of Congress.” *Database* 17, no. 2 (1994): 57-60.

Valls, C. H. “Multimedia in museums - an overview of its developments.” *Program-Automated Library and Information Systems* 28, no. 3 (1994): 263-274.

Wallace, J. “Project Chapman: The Direct Delivery of Digital Smithsonian Photographic Images Via the Internet.” *Visual Resources* 10, no. 1 (1994): 57-60.

Wang, J. Z., J. Li, D. Chan, and G. Wiederhold. “Semantics-sensitive Retrieval for Digital Picture Libraries.” *D-Lib Magazine* 5, no. 11, November (1999). www.dlib.org/dlib/november99/wang/11wang.html, (acc. Oct 2000).

Wayner, P. “Optical Character Recognition: an inside look at how optical character recognition works.” *Byte* December 18, no. 13 (1993): 203-120.

Waibel, G. *Produce, Publish and Preserve: A Holistic Approach to Digital Assets Management*. www.bampfa.berkeley.edu/moac/imaging/index.html, (acc. Oct 2000).

Weibel, S. L., and C. Lagoze. “An Element Set to Support Resource Discovery: The State of the Dublin Core.” *International Journal on Digital Libraries* 1, no. 2 (1997): 176-186.

Weibel, S., and E. Miller. “Image Description on the Internet.” *D-Lib Magazine* 3, no. 1, January (1997). www.dlib.org/dlib/january97/oclc/01weibel.html, (acc. Oct 2000).

Weibel, S., and J. Hakala. “DC-5: The Helsinki Metadata Workshop: A Report on the Workshop and Subsequent Developments.” *D-Lib Magazine* 4, no. 2, February (1998). www.dlib.org/dlib/february98/02weibel.html, (acc. Oct 2000).

Weibel, S., J. Kunze, C. Lagoze, and M. Wolf. “Dublin Core Metadata for Resource Discovery.” 1998, Internet Engineering Task Force, ftp://ftp.isi.edu/in-notes/rfc2413.txt, (acc. Oct 2000), (created Sept 1998).

Weinstein, W. “Designing an Image Database: A Holistic Approach.” *Visual Resources* 10, no. 1 (1994): 49-56.

Weise, J. “Image Database Access Architecture.” *(version I.I., draft I),* 1997, University of Michigan, images.umdl.umich.edu/info/arch/Image_Tech_Dev_Plan.html, (acc. Oct 2000), (created June 1997).

Weise, J. “Overview of Media Database Issues and Technologies.” *OIT Digital Media Solutions,* www.oit.itd.umich.edu/projects/DMI/answers/pickmediadb.html, (acc. Jan 2000).

Weise, J. P. “Data Transfer Time.” *Preparing Quality Images for Computer Networks,* University of Michigan, www-personal.umich.edu/~jweise/quality/DataTransferTime.HTML, (acc. Oct 2000).

Witten, I. H., A. Moffat, and T. C. Bell. *Managing gigabytes: Compressing and Indexing Documents and Images.*: Van Nostrand Reinhold, 1994.

Witten, I. H., C. Nevill-Manning, R. McNab, and S. J. Cunningham. “A Public Library Based on Full-text Retrieval.” *Communications of the ACM* 41, no. 4, April (1998): 71-75.

Woodward, H. “Selecting journals for digitisation: piecing together the puzzle to create a European model.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 77-83.

Wyatt, V., and G. MacLean. “Imaging Databases in Research and Training: Global Perspectives and New Research Technologies.” *Visual Resources* 10, no. 1 (1994): 41-48.

Zorich, D. M. *Introduction to Managing Digital Assets: Options for Cultural and Educational Organizations.* Los Angeles: Getty Information Institute, 1999.

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

10. Collaboration

"Guidelines and Standards: Getting Started." Colorado Digitization Project, coloradodigital.coalliance.org/standard.html, (acc. Jan 2000), (created 1999).

"Higher Education Digitisation Service (HEDS)." heds.herts.ac.uk/, (acc. Oct 2000).

"Internet Library of Early Journals (January 1996 - August 1998). A Project in the eLib Programme: Final Report." www.bodley.ox.ac.uk/ilej/papers/fr1999/fr1999.htm, (acc. Oct 2000), (created March 1999).

"Museums and the Online Archive of California." Online Archive of California, www.oac.cdlib.org/dynaweb/ead/moac/, (acc. Oct 2000).

The effects of digitisation on library and information services: proceedings of a conference organised by NORDINFO and the British Library (Research and Development Department), held at the Holiday Inn, Edinburgh, Scotland, 17-20 September 1992, *British Library R & D report 6098*. 1993, London: British Library Research and Development Department.

Barber, D. "OhioLINK: A Consortial Approach to Digital Library Management." *D-Lib Magazine* 3, no. 4, April (1997). www.dlib.org/dlib/april97/04barber.html, (acc. Oct 2000).

Bunker, G., and G. Zick. "Collaboration as a Key to Digital Library Development: High Performance Image Management at the University of Washington." *D-Lib Magazine* 5, no. 3, March (1999). www.dlib.org/dlib/march99/bunker/03bunker.html, (acc. Oct 2000).

Dempsey, L., T. Gardner, M. Day, and T. van der Werf. "International Information Gateway Collaboration: Report of the First IMesh Framework Workshop." *D-Lib Magazine* 5, no. 12, December (1999). www.dlib.org/dlib/december99/12dempsey.html, (acc. Oct 2000).

Endres, A. "Digital library and electronic publishing activities of German Professional societies." In *Proceedings of the 1997 IEEE Forum on Research and Technological Advances in Digital Libraries (May 7-9, 1997, Washington DC)*, 1997. 63-68.

Gill, T., C. Grout, and L. Smith. "Visual arts, museums and cultural heritage information standards: a domain-specific review of relevant standards for networked information discovery." Visual Arts Data Service, vads.ahds.ac.uk/training_advice/standards/standards_review/info%20standards.html, (acc. Oct 2000), (created March 1997).

Lossau, N., and S. Cramme. "DigiZeit (Digitisation of journals) - a joint effort of special subject collection libraries in Germany." In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark's Research Library, 2000. 19-23.

McKnight, C. "Many Projects that Depend on Collaboration." *Communications of the ACM* 41, no. 4, April (1998): 86-87.

Mintzer, F. C., L. E. Boyle, A. N. Cazes, B. S. Christian, S. C. Cox, F. P. Giordano, H. M. Gladney, J. C. Lee, M. L. Kelmanson, A. C. Lirani, K. A. Magerlein, A. M. B. Pavani, and F. Schiattarella. “Toward On-line, Worldwide Access to Vatican Library Materials.” *IBM Journal of Research and Development* 40, no. 2 (1996). www.research.ibm.com/journal/rd/mintz/mintzer.html, (acc. Oct 2000).

Odlyzko, A. “Competition and Cooperation: Libraries and Publishers in the Transition to Electronic Scholarly Journals (revised version April 17, 1999).” 1999, AT&T Labbs-Research, www.research.att.com/~amo/doc/comptition.cooperation.pdf, (acc. Oct 2000), (created April 1997).

Odlyzko, A. “Economics of Electronic Publishing: Journals Pricing and User Acceptance, The Economics of Electronic Journals.” *(paper presented at ARL’s Scholarly Communication and Technology Conference, Emory University, April 24-25, 1997),* 1997, Association of Research Libraries, arl.cni.org/scomm/scat/odlyzko.html, (acc. Oct 2000), (created July 1997).

Prescott, A. “The ‘Politics of Dancing’: Collaboration in Digital Projects.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 85-94.

Reid, N. H. “Photographic Archives: Aberdeen, Dundee and St Andrews.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 106-119.

Ross, S. “Consensus, Communication, and Collaboration: Fostering Multidisciplinary Cooperation in Electronic Records.” In *Proceedings of the DLM-Forum on Electronic Records, Brussels, 18-20 December 1996. INSAR: European Archives News, Supplement II*, Luxembourg: Office for Official Publications of the European Communities, 1997. 330-336.

Schutzer, D. “A Need For A Common Infrastructure: Digital Libraries and Electronic Commerce.” *D-Lib Magazine* 2, no. 4, April (1996). www.dlib.org/dlib/april96/04schutzer.html, (acc. Oct 2000).

Schwartz, W. “DIEPER - providing web access to retro-digitised periodicals at multiple sites.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 25-30.

Schwartz, W., and A. Regent. “DIEPER Project: Services Modelling.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 42-45.

Willett, P. “The Victorian Women Writers Project: The Library as a Creator and Publisher of Electronic Texts.” *The Public-Access Computer Systems Review,* 7, no. 6 (1996). http://info.lib.uh.edu/pr/v7/n6/will7n6.html, (acc. Oct 2000).

Willis, D., and D. Lathin. “Make Your Digital Imaging Program Successful: Involve the Stakeholders.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 7.

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

11. Case Studies

"About JSTOR." Journal Storage (JSTOR), www.jstor.ac.uk/about/need.html, (acc. Oct 2000), (created 2000).

"American Memory DTD for Historical Documents." Library of Congress, American Memory, memory.loc.gov/ammem/amdtd.html, (acc. Oct 2000), (created June 1998).

"AMICO: Art Museum Image Consortium home page." AMICO, www.amico.net, (acc. Oct 2000).

"Archiving and Distribution Working Group Charter." GII Virtual Roundtable, www.globalcollaboration.org/jtc1/igbt/adwg.html, (acc. Oct 2000), (created Nov 1999*).

"ArtsConnectEd." www.artsconnected.org, (acc. Oct 2000).

"Atlas Van der Hagen and Atlas Beudeker." Royal Library of the Netherlands, www.kb.nl/, (acc. Oct 2000).

"Bildarchiv zur Kunst und Architektur (The Image Archive for Art and Architecture)." University of Marburg, bildindex.de/intro.htm, (acc. Oct 2000).

"'Bildarvet Project', Digitalizing the Finish-Swedish heritage of pictures for research." Âbo Akademi University, www.abo.fi/library/bildarvet/welcomee.sht, (acc. Oct 2000), (created Oct 1998).

"Center for Retrospective Digitization, Göttingen State and University Library." Göttinger Digitalizierungs-Zentrum, www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html, (acc. Oct 2000), (created June 1999).

"COLLAGE-project by Guildhall Library." Corporation of London, collage.nhil.com/index.htm, (acc. Oct 2000), (created 1997).

"Conceptual and Intermedia Arts Online." www.bampfa.berkeley.edu/ciao, (acc. Oct 2000).

"Decision Matrix/Workflows." Oxford University, www.bodley.ox.ac.uk/scoping/matrix.htm, (acc. Oct 2000).

"Deutschen Kolonialgesellschaft (German Colonial Society)." Stadt-Und Universitätsbibliothek Frankfurt am Main, www.stub.uni-frankfurt.de/kolonial/kginfo.htm, (acc. Oct 2000), (created Jan 2000).

"Digital Camera Reviews." Digital Photography Reviews, www.inconference.com/digicam/camera.html, (acc. Oct 2000).

"Digital imaging and optical storage technology helps open up railway archives at National Railway Museum." *Electronic Library* 11, no. 2 (1993): 108-109.

"Digital Library Federation home page." www.clir.org/diglib, (acc. Oct 2000), (created Aug 2000).

"Digital Library Production Services (DLPS) at the University of Michigan." images.umdl.umich.edu, (acc. Oct 2000).

"Digital Library Program." The Library of Virginia, www.lva.lib.va.us/dlp, (acc. Oct 2000), (created 1995).

“Digitisation of the new Bibliotheque Nationale de France.” *Electronic Library* 12, no. 5, October (1994): 301-303.

“Digitising the Primary Source Material at the English Faculty Library, University of Oxford.” University of Oxford, info.ox.ac.uk/jtap/reports/digit/, (acc. Oct 2000), (created Oct 1997).

“Digitizing the Collection.” California Heritage Collection, sunsite.berkeley.edu/CalHeritage/digital.html, (acc. Oct 2000), (created Nov 1997).

“Documents about the DNER.” Joint Information Systems Committee, www.jisc.ac.uk/pub#dner, (acc. Oct 2000), (created Aug 2000).

“Ein neues Konvertierungsprogramm der SUB Göttingen/GDZ zur Umwandlung bibliographischer Pica/GBV-Daten in RDF/XML.” Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/gdz-tools/pica2xml/pica2xml_index_de.html, (acc. Oct 2000), (created 1997).

“Electronic Library Image Service for Europe Project (ELISE 1 and 2).” nile.dmu.ac.uk/elise, (acc. Oct 2000), (created Feb 2000*).

“Frequently Asked Questions.” *The Abilene Project,* www.ucaid.edu/abilene/html/faq-general.html, (acc. Oct 2000), (created 1997).

“Functional Requirements for Evidence in Recordkeeping.” University of Pittsburgh, School of Information Science, www.lis.pitt.edu/~nhprc, (acc. Oct 2000).

“Guidance on expressing the Dublin Core within the Resource Description Framework (RDF).” Dublin Core Metadata Initiative, www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/, (acc. Oct 2000), (created July 1999*).

“Guidelines and Standards: Getting Started.” Colorado Digitization Project, coloradodigital.coalliance.org/standard.html, (acc. Jan 2000), (created 1999).

“Higher Education Library Image Exchange project (HELIX).” www.helix.dmu.ac.uk/, (acc. Oct 2000), (created June 1998).

“Historic Pittsburgh, a joint project of the University of Pittsburgh and the Historical Society of Western Pennsylvania.” digital.library.pitt.edu/pittsburgh/, (acc. Oct 2000), (created Aug 2000*).

“Image Access homepage.” www.imageaccess.com, (acc. Aug 2000), (created 1999).

“Image Scanning.” Harvard University, Library Digital Initiative, hul.harvard.edu/ldi/html/image_scanning.html, (acc. Oct 2000), (created 1998).

“Images Collection: Introduction.” *Library and Information Centre,* The Royal Society of Chemistry, www.rsc.org/lic/imagesintro.htm, (acc. Oct 2000).

“Internet Library of Early Journals (January 1996 - August 1998). A Project in the eLib Programme: Final Report.” www.bodley.ox.ac.uk/ilej/papers/fr1999/fr1999.htm, (acc. Oct 2000), (created March 1999).

“JSTOR Usage Statistics.” Journal Storage, stats.jstor.org, (acc. Oct 2000), (created 2000).

“Lessons Learned: National Digital Library Competition.” The Library of Congress, lcweb2.loc.gov/ammem/award/lessons/lessons.html, (acc. Oct 2000), (created April 1999).

“Making of America II.” *The Making of America Testbed Project White Paper. Version 2.0, February 15,* 1998, http://sunsite.berkeley.edu/moa2/wp-v2.html, (acc. Oct 2000), (created March 1998).

“Manchester Computerised Local Image Collection.” Manchester Local Studies Unit, www.manchester.gov.uk/mccdlt/lsuimage/home.htm, (acc. Oct 2000), (created Aug 2000*).

“Marriage, Women, and the Law, 1815-1914: Studies in Scarlet.” Research Libraries Group, lyra.rlg.org/scarlet/index.html, (acc. Oct 2000), (created July 2000*).

“"New Library: the People’s Network": the Government’s response, Cm 3887.” 1998, HMSO, www.culture.gov.uk/new-library.htm, (acc. Oct 2000).

"PANDORA Project: Preserving and Accessing Networked Documentary Resources of Australia." National Library of Australia, www.nla.gov.au/pandora, (acc. Oct 2000), (created Jun 2000*).

"Planning Digital Projects for Historical Collections in New York State." New York Public Library, digital.nypl.org/brochure, (acc. Oct 2000), (created Aug 1999).

"REX: image database of the Danish Royal Library." www.kb.dk/formidl/rex/eng/index.htm, (acc. Oct 2000), (created Nov 1999*).

"Sample Scans: The Electronic Archive of Early American Fiction." University of Virginia Library, etext.lib.virginia.edu/projects/scantest.html, (acc. Oct 2000).

"Sample Structural Metadata for the Coolidge Image-Only Documents." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/timag-samp.html, (acc. Oct 2000), (created March 1997).

"Sample Structural Metadata for the Coolidge Photos." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/photo-samp.html, (acc. Oct 2000), (created March 1997).

"Sample Structural Metadata for the Coolidge SGML Encoded Documents." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/sgml-samp.html, (acc. Oct 2000), (created April 1997).

"Technical Knowledge Centre & Library of Denmark Revolutionises the Library World." Technical Knowledge Centre & Library of Denmark, www.dtv.dk/new/press/210199_e.htm, (acc. Oct 2000), (created Dec 1999*).

"The Aberdeen Bestiary: A pilot scheme for producing digitised images of manuscripts on the World Wide Web." Aberdeen University, www.clues.abdn.ac.uk:8080/besttest/firstpag.html, (acc. Oct 2000).

"The National Digital Library." *Information Retrieval & Library Automation* 30, no. 5, October (1994): 1-3.

"The Pandora Logical Model, Version 2." National Library of Australia, www.nla.gov.au/pandora/ldmv2.html, (acc. Oct 2000), (created Nov 1997).

"Towards a Learning Nation: the Digital Contribution Recommendations Proposed by the Federal Task Force on Digitization, Final Report, 1997." National Library of Canada, www.nlc-bnc.ca/coopprog/finalreport/eindex.htm, (acc. Oct 2000), (created Dec 1997).

"VASARI: Visual Arts System for Archiving and Retrieval of Images." University of Southampton, www.ecs.soton.ac.uk/~km/projs/vasari.html, (acc. Oct 2000).

Arcoli, A., and B. Washburn. "Delivering the AMICO Library™ through RLGs Eureka® Service." In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 128-129.

Arms, C. "Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress." *D-Lib Magazine* 2, no. 5, May (1996). www.dlib.org/dlib/may96/loc/05c-arms.html, (acc. Oct 2000).

Arms, C. R. "Keeping Memory Alive: Practices for Preserving Digital: Content at the National Digital Library Program of the Library of Congress." *RLG DigiNews* 4, no. 3, June 15 (2000). www.rlg.org/preserv/diginews/diginews4-3.html#feature1, (acc. Oct 2000).

Arnott, M., I. Beavan, and J. Milne. "The On-Line Bestiary Project." In *ELVIR Conference, 3rd, De Montfort University, Milton Keynes [Proceedings]*, London: ASLIB, 1996.

Atkins, D. E. "The University of Michigan Digital Library Project: The Testbed." *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/07atkins.html, (acc. Oct 2000).

Bagnall, R. "Advanced Papyrological Information System." *A Joint Project of Columbia University, Duke University, Princeton University, The University of California, Berkeley, The University of Michigan, Université Libre de Bruxelles, Yale University: Narrative,* odyssey.lib.duke.edu/papyrus/texts/APISgrant.html, (acc. Oct 2000).

Bagnall, R. S. *Digital Imaging of Papyri: a report to the Commission on Preservation and Access.* Washington,

D.C.: The Commission on Preservation and Access, 1995.

Baker, C. “A marriage of high-tech and fine art: the National Gallery’s Micro Gallery project.” *Program* 27, no. 4, October (1993): 341-352.

Barisano, E. *Télédetection et cartographie thématique en archéologie.* Paris: CRN, 1988.

Bennett, E. Z. “Digitizing Science: JSTOR Faces New Challenges.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 26-27.

Berger, B. “Digital Imaging Projects at Cornell University.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 49-57.

Bescós Ramón, J., F. Jaque, and L. Montoto. “Reflectance and Optical Contrast of Old Manuscripts: Wavelength Dependance.” In *Scanning imaging: ECO1: 21-23 September 1988*, edited by Tony Wilson, 1989. 258-62.

Besser, H., and M. Snow. “Access to Diverse Collection in a University Setting: The Berkeley Dilemma.” In *Beyond the Book: Extending MARC for Subject Access to Non-Traditional Materials*, edited by Pat Moholt and Toni Petersen, New York: G. K. Hall & Co, 1990.

Bevan, I. “Beasts on the Screen: The Digitisation of the Aberdeen Bestiary: A Case Study.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 58-64.

Bishop, A. P. “Working Towards an Understanding of Digital Library Use: A Report on the User Research Efforts of the NSF/ARPA/NASA DLI Projects.” *D-Lib Magazine* 1, no. 10, October (1995). www.dlib.org/dlib/october95/10bishop.html, (acc. Oct 2000).

Bonn, M. “Building a Digital Library: The Stories of the Making of America.” University of Michigan, www.umdl.umich.edu/dlps/mbonn-saunders.html, (acc. Oct 2000).

Bonn, M. S., J. K. MacKie-Mason, W. P. Lougee, and J. F. Riveros. “A Report on the PEAK Experiment.” *D-Lib Magazine* 5, no. 6, June (1999). www.dlib.org/dlib/june99/06bonn.html, (acc. Oct 2000).

Bot, M., J. Burgemeester, and H. Roes. “The Cost of Publishing an Electronic Journal: A general model and a case study.” *D-Lib Magazine* 4, no. 11, November (1998). www.dlib.org/dlib/november98/11roes.html, (acc. Oct 2000).

Bouché, N. *Digitization for scholarly use: The Boswell Papers Project at the Beinecke Rare Book and Manuscript Library.* Washington, DC: Council on Library and Information Resources, 1999.

Bowen, W. G. “JSTOR and the Economics of Scholarly Communication.” *The Andrew W. Mellon Foundation, Journal Storage Project,* 1995, www.mellon.org/jsesc.html, (acc. Oct 2000), (created Oct 1995).

Bramich, K., and J. Cannon. *Capturing the Big Picture: Strategies for Image Libraries at the National Museum of Australia.* abridged version ed.: National Museum of Australia, 1998. www.nma.gov.au/aboutus/products/bigpic.pdf, (acc. Oct 2000).

Breu, M., A. Brüggermann-Klein, C. Haber, and R. Weber. “The MeDoc distributed electronic library - Accounting and security aspects.” In *Proceedings of the ICCC/IFIP Conference (April 1997, Canterbury, England)*, Canterbury: University of Kent, 1997.

Butter, K., R. Chandler, and J. Kunze. “The Cigarette Papers: Issues in Publishing Materials in Multiple Formats.” *D-Lib Magazine* 2, no. 11, November (1996). www.dlib.org/dlib/november96/11butter.html, (acc. Oct 2000).

California Heritage Collection. “Digitizing the Collection: Overview.” University of California, Berkeley, The Bancroft Library, sunsite.berkeley.edu/CalHeritage/digital.html, (acc. Oct 2000), (created Nov 1997*).

Campbell, J. “Archives, Imaging and CD-ROMs: case studies of projects at United Distillers and Prudential.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 66-71.

Carey, P. “PIX Project.” In *Towards the Digital Library: The Initiatives for Access Programme*, edited by L Carpenter, S Shaw and A Prescott, London: British Library, 1998.

Carnaffin, E. “Gateshead 1900, a CD-I guide.” *The Local Studies Historian* 14, no. 2, Winter (1995).

Cathaly, G. “Mass Digitisation Production Chain at the Bibliotheque Nationale de France.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 15-19.

Chapman, S., P. Conway, and A. R. Kenney. “Digital Imaging and Preservation Microfilm: The Future of the Hybrid Approach for the Preservation of Brittle Books.” CLIR, Commission on Preservation and Access, www.clir.org/cpa/archives/hybridintro.html, (acc. Oct 2000).

Chestnutt, D. “The Model Editions Partnership: Historical Editions in the Digital Age.” *D-Lib Magazine* 1, no. 11, November (1995). www.dlib.org/dlib/november95/11chesnutt.html, (acc. Oct 2000).

Chestnutt, D. R. “The Model Editions Partnership: "Smart Text" and Beyond.” *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/07chesnutt.html, (acc. Oct 2000).

Coleman, R. “Digitzing the Record of a Colonial Culture: Ferguson 1840-45.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 16.

Colet, L. S., K. Keller, and E. Landsberg. “Digitizing Photographic Collections: A Case Study at the Museum of Modern Art, NY.” *Spectra* 25, no. 2 (1997): 22-27.

Committee on Electronic Information (CEI) - Content Working Group. “An Integrated Information Environment for Higher Education: Developing the Distributed, National Electronic Resource.” *JISC (Joint Information Systems Committee),* www.jisc.ac.uk/cei/dner_colpol.html, (acc. Oct 2000), (created Dec 1997).

Commonwealth Higher Education Management Service. “Evaluation of the UK Pilot Site Licence Initiative - Phase II.” *HEFCE Ref 98/22,* 1998, National Information Services and Systems (NISS), www.niss.ac.uk/education/hefce/pub98/98_22.html, (acc. Oct 2000), (created May 1998).

Conway, P. “Selecting Microfilm for Digital Preservation: A Case Study from Project Open Book.” *Library Resources and Technical Services* 40, no. 1 (1996a): 67-77.

Conway, P. “Yale University Library’s Project Open Book.” *D-Lib Magazine* 2, no. 2, February (1996). www.dlib.org/dlib/february96/yale/02conway.html, (acc. Oct 2000).

Conway, P. *Preservation in the Digital World.* Washington, D.C.: Commission on Preservation and Access, 1996b.

Conway, P. *Selecting Microfilm for Digital Preservation: A Demonstration Project: Performance Report on the production conversion phase of Project Open Book.* New Haven, Conn.: Yale University Library, 1996a.

Coyle, K. “The Virtual Union Catalog: A Comparative Study.” *D-Lib Magazine* 6, no. 3, March (2000). www.dlib.org/dlib/march00/coyle/03coyle.html, (acc. Oct 2000).

Crane, G. “The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology.” *D-Lib Magazine* 4, no. 1, January (1998). www.dlib.org/dlib/january98/01crane.html, (acc. Oct 2000).

D’Amato, D. P., and R. C. Klopfenstein. “Requirements and Options for the Digitization of the Illustration Collections of the Natural Museum of Natural History.” 1996, www.nmnh.si.edu/cris/techrpts/imagopts, (acc. Oct 2000), (created March 1996).

Danish National Library of Science and Medicine. “Flora Danica Online.” www.pictures.dnlb.dk/Homepage/welcome.html, (acc. Oct 2000).

De Gennaro, R. “JSTOR: Building an Internet Accessible Digital Archive of Retrospective Journals.” *(63rd IFLA General Conference, Conference Programme and Proceedings, August 31 - September 5, 1997),* 1997, IFLANET, ifla.inist.fr/IV/ifla63/63genr.htm, (acc. Oct 2000), (created Sept 1997).

Duggan, H. N. “Creating an Electronic Archive of *Piers Plowman*.” University of Virginia,

www.village.virginia.edu/piers/report94.html, (acc. Aug 2000), (created 1994).

Duggan, M. K. “19th-Century California Sheet Music.” University of California, Berkeley, www.sims.berkeley.edu/~mkduggan/project.html, (acc. Oct 2000).

Eaves, M., R. N. Essick, and J. V. (eds). “The William Blake Archive.” Virginia University, www.iath.virginia.edu/blake, (acc. Oct 2000), (created 1996).

Eck, R. “The digitization program of travel books and early North-Americana at the State and University of Lower Saxony (SUB) in Göttingen.” Göttinger DigitalisierungsZentrum (GDZ), www.sub.uni-goettingen.de/gdz/en/projects/itinnorda_en.html, (acc. Oct 2000), (created April 1998).

Enders, M. “E-DOC format for DIEPER. 2nd edition.” www.sub.uni-goettingen.de/gdz/dieper, , (created *forthcoming*).

Endres, A. “Digital library and electronic publishing activities of German Professional societies.” In *Proceedings of the 1997 IEEE Forum on Research and Technological Advances in Digital Libraries (May 7-9, 1997, Washington DC)*, 1997. 63-68.

Ester, M. “Background.” In *Introduction, Frank Lloyd Wright Presentation and Conceptual Drawings*, Venice, CA: Luna Imaging Inc., 1995.

Eyre, J. “ELISE II: Networked Delivery of Images and User Interface Issues.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 31-34.

Federal Bureau of Investigation: Criminal Justice Information Service. *Electronic Fingerprint Transmission Specification (EFTS).* Washington, DC: CJIS, 1999.

Fenton, E., K. Guthrie, and A. Kirchhoff. “Digitising journals: Highlights from JSTOR’s experience.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 31-41.

Fleischhauer, C. “Digital Historical Collections: Types, Elements, and Construction.” 1996, American Memory: Historical Collections for the National Digital Library, memory.loc.gov/ammem/elements.html, (acc. Oct 2000), (created Aug 1996).

Fleischhauer, C. “Selecting Collections and Selecting Technology: American Memory at the Library of Congress.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 20.

Fowler, M. “High-Resolution Satellite Imagery in Application - A Russian Satellite Photograph of the Stonehenge Region.” *Antiquity* 70 (1996): 667-671.

Fresko, M. “Results of the Comparative Study of Digital Preservation Guidelines.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 46-51.

Frew, J., M. Freeston, R. B. Kemp, J. Simpson, T. Smith, A. Wells, and Q. Zheng. “The Alexandria Digital Library Testbed.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/alexandria/07frew.html, (acc. Oct 2000).

Gagos, T., with, P. Daub, A. Loftus, and S. Zachary. “Scanning the Past: A Modern Approach to Ancient Culture.” *Library Hi Tech* 14, no. 1 (1996): 11-22.

German Research Foundation, D. *Retrospektive Digitaliserung von Bibliotheksbeständen: Berichte der von der Deutschen Forschunsgemeinschaft einberufenen Facharbeitsgruppen "Inhalt" und "Technik".* Berlin: DBI, 1998. www.sub.uni-goettingen.de/java_home-e.htm, (acc. Oct 2000).

Getz, M. “Questionnaire for Digital Scriptorium Users.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSquery.pdf, (acc. Oct 2000), (created March 1998).

Getz, M. “The Digital Scriptorium: Medieval Manuscripts on the World Wide Web.” Vanderbilt University, www.vanderbilt.edu/Econ/faculty/Getz/DSvision.pdf, (acc. Oct 2000), (created March 1998).

González García, P. “Computerization Project for the Archivo General de Indias.” In *International Conference: Archiving and Disseminating Historical Machine-readable Data*, Leiden, Netherlands, 1992.

González García, P. “El Archivo General de Indias y su Proyecto de Informatización. Neuvas Posibilidades Para la Investigación.” *Cuardernos de Historia Moderna* 15 (1994): 231-49.

González García, P. “Historical Documentation and Digital Conversion of Images at the Proyecto de Infomatización of the Archivo de Indias, Sevilla.” *Microform Review* 18, no. 4 (1989): 217-21.

González García, P. “New Technology and the Reconstruction of the Archival Heritage.” In *Proceedings of the XXX International Conference of the Round Table on Archives, Thessaloniki, 1994*, Dordrecht, The Netherlands: International Council on Archives, 1998. 125-29.

González García, P. “Novas Tecnologias no Arquivo Geral da Indias.” *Acervo, Revista do Arquivo Nacional de Rio de Janeiro* 7, no. 1-2 (1994): 75-89.

González García, P. “Proyecto de Informatización del Archivo General de Indias.” In *Actas del II Congreso Iberoamericano de Informática y Documentación. Mar del Plata (Argentina 1998)*, Madrid: Centro Regional para la Enseñanza de la Informática, 1989. 409-12.

González García, P. “Proyecto de Informatización del Archivo General de Indias.” In *Actas de las I Jornadas de Archivística de Euskadi, 1990. IRARGI no. 3*, Vitorai, Spain: IRARGI, 1990. 261-81.

González García, P. *Computerization of the Archivo General de Indias: Strategies and Results.* Washington, DC: Council on Library and Information Resources, 1998. www.clir.org/pubs/reports/gonzalez/contents.html, (acc. Oct 2000).

Gordon, C. “Pattern of User Queries in an ICONCLASS Database.” *Visual Resources* XII (1996): 177-186.

Gould, C. C. *Information Needs in the Humanities: An Assessment.* Stanford, CA: The Research Libraries Group, 1988.

Grout, C. “COLDIP: The Corporation of London Data Imaging Project. One of a New Age of Imaging Projects.” Visual Arts Data Service, http://vads.ahds.ac.uk/training_advice/case_studies/Coldip.html, (acc. Oct 2000), (created Nov 1997).

Guthrie, K. M. “JSTOR: From Project to Independent Organization.” *D-Lib Magazine* 3, no. 7/8, July/August (1997). www.dlib.org/dlib/july97/07guthrie.html, (acc. Oct 2000).

Guthrie, K. M. “JSTOR: The Development of a Cost-Driven, Value Based Pricing Model.” *(paper presented at ARL’s Scholarly Communication and Technology Conference, Emory University, April 24-25, 1997),* 1997, Association of Research Libraries, arl.cni.org/scomm/scat/guthries.html, (acc. Oct 2000), (created April 1997).

Harris, O. “United we stand - Croyden’s multidisciplinary historical database.” *IT’s News, the Journal of the Library Association Information Technology Group* 33, no. June (1996).

Harum, S. L., W. H. Mischo, and B. R. Schatz. “Federating Repositories of Scientific Literature: An Update on the Digital Library Initiative at the University of Illinois at Urbana-Champaign.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/07harum.html, (acc. Oct 2000).

Holm, J., and N. Judge. “Electronic Photography at the NASA Langley Research Center.” In *Proceedings, IS&T 48th Annual Conference*, Washington, DC, 1995. 436-441.

Holt, B., and L. Hartwick. “Retrieving Art Images by Image Content: The UC Davis QBIC Project.” In *Electronic Library and Visual Information Research, ELVIRA 1: Proceedings of the First ELVIRA Conference, Held in May 1994 at De Montfort University, Milton Keynes*, edited by Mel Collier and Kathryn Arnold, London: Aslib, 1995. 93-100.

Hopkin, D. “Shifting the Focus: Digital Imaging & Photographic Collections Management at the National Railway Museum.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 72-80.

Hughes, C. A. “Lessons Learned: Digitization of Special Collections at The University of Iowa Libraries.” *D-Lib Magazine* 6, no. 6, June (2000). www.dlib.org/dlib/june00/hughes/06hughes.html, (acc. Oct 2000).

Indiana University Electronic Records Project. “A Methodology for Evaluating Recordkeeping Systems.” *Draft (March),* 1997, Indiana University, www.indiana.edu/~libarche/97method.html, (acc. Oct 2000), (created April 1998*).

JSTOR. “Moving Wall.” Journal Storage, www.jstor.org/journals/movingwall.html, (acc. Oct 2000).

Jupp, B. “The Internet Library of Early Journals.” *Aslib Proceedings* 49, no. 6 (1998): 153-158.

Kenney, A. R. “Digital to Microfilm Conversion: A Demonstration Project, 1994-1996, Final Report.” 1997, Cornell University Library, Ithaca, NY, www.library.cornell.edu/preservation/com/comfin.html, (acc. Oct 2000).

Kenney, A. R. “Digital-to-Microfilm Conversion: An Interim Preservation Solution.” *Library Resources & Technical Services* 37, no. October (1993): 380-401.

Kenney, A. R. “The Cornell Digital Microfilm Conversion Report: Final Project to NEH.” *RLG DigiNews* 1.2, no. August (1997). www.rlg.org/preserv/diginews/diginews2.html#com, (acc. Oct 2000).

Kessler, J. “The French Minitel: Is There Digital Life Outside of the "US ASCII" Internet? A Challenge or Convergence?” *D-Lib Magazine* 1, no. 11, December (1995). www.dlib.org/dlib/december95/12kessler.html, (acc. Oct 2000).

Kiernan, K. “Digital Preservation, restoration, and the dissemination of medieval manuscripts.” In *Gateways, gatekeepers, and roles in the information omniverse: proceedings of the third symposium: November 13-15, 1993*, edited by A Okerson and D Mogge, Washington, DC, 1994.

Kiernan, K. S. “The Electronic Beowulf.” *Computers in Libraries* 15, no. 2, February (1995): 14-15.

Kingma, B. R. “The Costs of Print, Fiche, and Digital Access: The Early Canadiana Online Project.” *D-Lib Magazine* 6, no. 2, February (2000). www.dlib.org/dlib/february00/kingma/02kingma.html, (acc. Oct 2000).

Klak, J. “Pretty pixels: digital imaging in Kirkcaldy.” *The Local Studies Librarian* 16, no. 2(3) (1997): 1-9.

Klavans, J. L. “New Center at Columbia University for Digital Library Research: Fostering Interdisciplinary Research and Bridging Cultural Clashes.” *D-Lib Magazine* 2, no. 3, March (1996). www.dlib.org/dlib/march96/klavans/03klavans.html, (acc. Oct 2000).

Knoll, A. “Digitisation of Rare Library Materials: Compound Documents - Manuscripts and Old Printed Books.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 23-25.

LC21: A Digital Strategy for the Library of Congress. Washington, DC: National Academy Press, 2000. http://www.nap.edu/books/0309071445/html/, (acc. Oct 2000).

Leggate, P. “Internet Library of Early Journals Annual Report.” 1997, Oxford University, www.bodley.ox.ac.uk/ilej/papers/ar1997.htm, (acc. Oct 2000), (created Aug 1997).

Leggate, P. “Internet Library of Early Journals Project.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 20-22.

Library of Congress: National Digital Library Program and the Conservation Division. “Conservation Implications of Digitization Projects.” Library of Congress, American Memory, memory.loc.gov/ammem/techdocs/conservation.html, (acc. Oct 2000), (created Sept 1999).

Lossau, N. “Digitisation - technical issues: Production at the Göttingen Digitisation Center (GDZ).” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 47-51.

Lossau, N., and F. Klaproth. “Digitization Efforts at the Center for Retrospective Digitization, Göttingen University Library.” *RLG DigiNews* 3, no. 1, February 15 (1999). www.rlg.org/preserv/diginews/diginews3-1.html#feature2, (acc. Oct 2000).

Lossau, N., and S. Cramme. “DigiZeit (Digitisation of journals) - a joint effort of special subject collection libraries in Germany.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 19-23.

MacKenzie, G. “Digitizing the Scottish Wills.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 139-149.

Martinez, K. “High Resolution Digital Imaging of Paintings: The VASARI Project.” *Microcomputers for Information Management* 8, no. 4, December (1991): 277-83.

McClung, P., (ed.) *RLG Digital Image Access Project*. 1995, Mountain View, CA: The Research Libraries Group.

Mecocci, A. “Practical Experiences and Techniques Used for Acquiring Images of Large Historical Buildings and Cultural Items.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 13-14.

Melrose, E. “All Things Bright and Beautiful? - The Public Librarian’s Perspective.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 95-101.

Mintzer, F. C., L. E. Boyle, A. N. Cazes, B. S. Christian, S. C. Cox, F. P. Giordano, H. M. Gladney, J. C. Lee, M. L. Kelmanson, A. C. Lirani, K. A. Magerlein, A. M. B. Pavani, and F. Schiattarella. “Toward On-line, Worldwide Access to Vatican Library Materials.” *IBM Journal of Research and Development* 40, no. 2 (1996). www.research.ibm.com/journal/rd/mintz/mintzer.html, (acc. Oct 2000).

Morrison, I. “SCRAN Basic Record Standard.” 1997, Scottish Cultural Resource Access Network (SCRAN), Edinburgh, www.scran.ac.uk/articles/scranbds.htm, (acc. Oct 2000).

Murdoch, J. W. “Aberdeen Art Gallery image database project: a prototype low-cost art image database.” *Art Libraries Journal* 21, no. 1 (1995): 23-26.

Nartker, T. A., and S. V. Rice. “OCR Accuracy: UNLV’s Third Annual Test.” *INFORM* 8, no. 8 (1994): 30-36.

National Council on Archives. *British Archives: The Way Forward.* London: National Council on Archives, 1998.

Niblack, W., and M. Flickner. “Find me the Pictures that Look Like This: IBM’s Image Query Project.” *Advanced Imaging* April (1993): 32-35.

Nichols, M. M. “Art and artifact: the digital option, a case study of the "Il Papa" manuscript.” In *Society of Dance History Scholars (U.S.). Conference (20th, 1997, Barnard College). Proceedings*, 1997. 187-190.

Noordermeer, T. “Deposit for Dutch Electronic Publications: Research and Practice in The Netherlands.” In *Research and Advanced Technology for Digital Libraries: First European Conference, ECDL ’97, Pisa, Italy, September 1-3, 1997*, edited by Carol Peters and Costantino Thanos, Berlin; New York: Springer, 1997.

North Carolina, General Assembly, Legislative Research Commission, and Committee on Digitization of Public Records by State Archives. *Digitization of public records by state archives: report to the 2000 session of the 1999 General Assembly of North Carolina.* Raleigh, NC: The Commission, 2000.

Nurminen, A. “Recording, Processing, and Archiving Carbonised Papyri.” 1995, Helsinki University of Technology, www.cs.hut.fi/papyrus/TOC.html, (acc. Oct 2000).

Oddos, J.-P. “La Numérisation des Archives Générales des Indes à Séville.” *Bulletin de la Bibliothèque de France* 39, no. 4 (1994): 82-86.

Oldervoll, J. “The Census of Norway from the Year 1801.” Dept. of History, University of Bergen, digitalarkivet.uib.no/index-eng.htm, (acc. Oct 2000).

Ormes, S. “A directory of UK public library networking projects.” In *The Internet, Networking and the Public Library*, edited by Sarah Ormes and Lorcan Dempsey: Library Association Publishing, published in association with UKOLN, 1997.

Ormes, S. “An overview of UK public library networking projects.” In *The Internet, Networking and the Public Library*, edited by Sarah Ormes and Lorcan Dempsey: Library Association Publishing, published in association with UKOLN, 1997.

Ormes, S. “Internet activity in public libraries.” *Ariadne* 3, no. May 1996 (1996). www.ariadne.ac.uk/issue3/libraries, (acc. Oct 2000).

Ormes, S., and L. Dempsey, (eds.) *The Internet, Networking and the Public Library*. 1997: Library Association Publishing, published in association with UKOLN.

Plante, R. L. “The NCSA Astronomy Digital Image Library. The Challenges of the Scientific Data Library.” *D-Lib Magazine* 3, no. 10, October (1997). www.dlib.org/dlib/october97/adil/10plante.html, (acc. Oct 2000).

Plotkin, R. C., and M. S. Schwartz. “Data Modeling for News Clip Archive: A Prototype Solution.” *D-Lib Magazine* 3, no. 5, May (1997). www.dlib.org/dlib/may97/plotkin/05plotkin.html, (acc. Oct 2000).

Prescott, A. “Constructing Electronic Beowulf.” In *Towards the Digital Library: The British Library’s ‘Initiatives for Access’ Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott, London: The British Library, 1998. 30-49.

Prescott, A. “The electronic Beowulf and digital restoration.” *Literary and Linguistic Computing* 12, no. 3 (1997): 185-95.

Pretzer, C. A., and C. M. Watts. “The Establishment of a Visual Database for Multimedia Use at Kansas State University.” *Visual Resources* X (1995): 349-356.

Rees, J. “Interactive Multimedia - The Brancusi Project: An Exploration.” *Information Services and Use* 13, no. 4 (1993): 297-302.

Reid, G. “The digitisation of heritage material: arguing for an interpretative approach based on the experience of the Powys Digital History Project.” *Program - Electronic Library and Information Systems* 34, no. 2, April (2000): 143-158.

Royan, B. “Project Management and Quality Assurance.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 28-37.

Royan, B. “SCRAN: a taste of Scotland and food for thought.” *Ariadne* 7 (1997). www.ariadne.ac.uk/issue7/scran, (acc. Oct 2000).

Rusch-Feja, D., and H. J. Becker. “Global Info: The German Digital Libraries Project.” *D-Lib Magazine* 5, no. 4, April (1999). www.dlib.org/dlib/april99/04rusch-feja.html, (acc. Oct 2000).

Rusch-Feja, D., and U. Siebeky. “Evaluation of Usage and Acceptance of Electronic Journals: Results of an Electronic Survey of Max Planck Society Researchers including Usage Statistics from Elsevier, Springer and Academic Press.” *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/rusch-feja/10rusch-feja-summary.html, (acc. Oct 2000).

Rütimann, H., and S. Lynn. *Computerization Project for the Archivo General de Indias, Seville, Spain. A Report to the Commission on Preservation and Access.* Washington, DC: Commission on Preservation and Access, 1992.

Schupbach, W. “The Iconographic Collections Videodisc at the Wellcome Institute for the History of Medicine, London.” *Program* 28, no. 3, July (1994): 279-285.

Schwartz, W. “DIEPER - providing web access to retro-digitised periodicals at multiple sites.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 25-30.

Schwartz, W., and A. Regent. “DIEPER Project: Services Modelling.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 42-45.

Seçilla, J., et al. *The Documents of the New World. Minsiterio de Cultura de España, Fundacián Ramá Areces, IBM España*. Madrid: Archivos y Bibliotecas. CD-ROM.

Sharp, L. H., M. D. Ott, and C. Fleishhauer. “Library of Congress Manuscript Digitisation Demonstration Project, Final Report.” 1998, Library of Congress, memory.loc.gov/ammem/pictel/pictel.pdf, (acc. Oct 2000), (created Oct 1998).

Small, L. M. “From paper to pixels (The SIL digitization program).” *Smithsonian* 31, no. 2, May (2000): 18.

Smith, A. *The Future of the Past: Preservation in American Research Libraries.* Washington, DC: Council on

Library and Information Resources (CLIR), 1999. www.clir.org/cpa/reports/pub82/pub82text.html, (acc. Oct 2000).

Starr, S. S. “Building the Collections of the California Digital Library.” *Issues in Science and Technology Librarianship* Winter (1998). http://www.library.ucsb.edu/istl/98-winter/article2.html, (acc. Oct 2000).

Stephenson, C., and P. McClung, (eds.) *Images Online: Perspectives on the Museum Educational Site Licensing Project*. Vol. 2, *The Museum Educational Site Licensing Project*. 1998, Los Angeles: The Getty Information Institute.

Stokes, J. R. “Imaging Pictorial Collections at the Library of Congress.” *RLG DigiNews* 3, no. 2, April 15 (1999). www.rlg.org/preserv/diginews/diginews3-2.html#feature, (acc. Oct 2000).

Stokes, J. R. “Sheet Music Scanning Procedures.” Historic American Sheet Music, Duke University, Rare Book, Manuscript and Special Collections Library, http://scriptorium.lib.duke.edu/sheetmusic/procedure.html, (acc. Oct 2000).

Süsstrunk, S. “Image Production Systems at Corbis Corporation.” *RLG DigiNews* 2, no. 4, August 15 (1998). www.rlg.org/preserv/diginews/diginews2-4.html#technical, (acc. Oct 2000).

Swain, D. J. “Enhancing and reconstructing digitized handwriting.” MSc, Virginia Polytechnic Institute and State University, 2000.

Tanner, S., and B. Robinson. “JISC Image Digitisation Initiative (JIDI) Feasibility Study (Final Report).” 1997, Higher Education Digitisation Service, Hatfield, heds.herts.ac.uk/resources/papers/jidi_fs.html, (acc. Oct 2000), (created July 1998).

The effects of digitisation on library and information services: proceedings of a conference organised by NORDINFO and the British Library (Research and Development Department), held at the Holiday Inn, Edinburgh, Scotland, 17-20 September 1992, *British Library R & D report 6098*. 1993, London: British Library Research and Development Department.

The European Visual Archive. “Proposal: European strategic project to stimulate the use of multimedia information content exploiting resources in the public archives.” 2000, European Visual Archive, www.eva-eu.org/propsal11.pdf, (acc. Oct 2000).

The Library of Congress/Ameritech. “Lessons Learned: National Digital Library Competition.” Library of Congress, memory.loc.gov/ammem/award/leesons/lessons.html, (acc. Oct 2000), (created April 1999).

Thorp, N., and G. Cannon. “The Electronic Edition of the Correspondence of James McNeill Whistler.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 96-105.

Turnbull, G., and A. Buchanan. “SCRAN Project Guidelines.” 1997, Scottish Cultural Resource Access Network, Edinburgh, www.scran.ac.uk/articles/projpaper.htm, (acc. Oct 2000), (created May 1997*).

Twycross, M., P. King, and A. Prescott. “The York Doomsday Project.” In *Towards the Digital Library: The British Library’s Initiatives for Access Programme*, edited by L Carpenter, S Shaw and A Prescott, London: British Library, 1998.

Valauskas, E. J. “Digital Images over the Internet: Rome Reborn at the Library of Congress.” *Database* 17, no. 2 (1994): 57-60.

Vicars-Harris, O. “Picturing Guildhall’s Treasures COLDIP: The Corporation of London Data Imaging Project.” *Computers and the History of Art* 6, no. 1 (1996).

Wallace, J. “Project Chapman: The Direct Delivery of Digital Smithsonian Photographic Images Via the Internet.” *Visual Resources* 10, no. 1 (1994): 57-60.

Watson, I. “The Durham Record.” *The Local Studies Librarian* 15, no. 2 (31) (1996): 2-6.

Watson, I. *The Durham Record.*, *Public Library Development Incentive Scheme (PLDIS) report no. 47*: British Library, 1996.

Webb, C. “The Ferguson Project: A Hybrid Approach to Reformatting Rare Australiana.” 1996, National Library

of Australia, www.nla.gov.au/nla/staffpaper/cwebb1.html, (acc. Oct 2000), (created Oct 1996).

Weinstein, W. “Designing an Image Database: A Holistic Approach.” *Visual Resources* 10, no. 1 (1994): 49-56.

Whitfield, S. “A Database for Cataloguing Chinese and Central Asian Manuscripts: The International Dunhuang Project.” In *Towards the Digital Library: The British Library’s ‘Initiatives for Access’ Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott, London: The British Library, 1998. 166-177.

Wiering, F., and H. Mulder, (eds.) *Digitisation of European Cultural Heritage*. 2000, Utrecht: Netherlands Utrecht University.

Willett, P. “The Victorian Women Writers Project: The Library as a Creator and Publisher of Electronic Texts.” *The Public-Access Computer Systems Review,* 7, no. 6 (1996). http://info.lib.uh.edu/pr/v7/n6/will7n6.html, (acc. Oct 2000).

Wise, A. “Developing the DNER for Learning and Teaching, 5 November 1999.” Joint Information Systems Committee, www.jisc.ac.uk/pub99/c05_99.html, (acc. Oct 2000), (created Nov 1999).

Wissenburg, A. “Digitising journals and the eLib programme.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 11-17.

Witten, I. H., S. J. Cunningham, and M. D. Apperley. “The New Zealand digital library project.” *D-Lib Magazine* 2, no. 11, November (1996). www.dlib.org/dlib/november96/newzealand/11witten.html, (acc. Oct 2000).

Yokley, J., J. Jansson, and D. A. Harmer. *Retrieving America’s past: digitizing 150 years of the Chicago*. Alexandria, VA: Audio Transcripts, Ltd, 2000. Audio Cassette.

Zillhardt, S. “Bibliotheca Universalis Assessment.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 28-30.

Zillhardt, S. “Bibliotheca universalis: a G-7 global information society pilot project.” *Paper presented at the IFLA General Conference, August 16-21 1998, Amsterdam. Paper 031-98-E,* 1998, www.ifla.org/IV/ifla64/031-98e.htm, (acc. Oct 2000), (created Aug 1998).

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

12. General

"An Introduction to Photos. 19th and Early 20th Century Photographic Processes." 1996, Public Record Office Kew, www.pro.gov.uk/preservation/photos.htm, (acc. Oct 2000), (created July 1999*).

"Decision Matrix/Workflows." Oxford University, www.bodley.ox.ac.uk/scoping/matrix.htm, (acc. Oct 2000).

"Doculabs Inc. home page." www.doculabs.com, (acc. Jan 2000).

"Electronic Imaging and Museums." *Visual Resources* 7 (1991).

"European Commission on Preservation and Access (ECPA)." www.knaw.nl/ecpa/, (acc. Oct 2000).

"European Visual Archives project (EVA)." www.eva-eu.org, (acc. Feb 2000), (created 1998).

"Higher Education Digitisation Service (HEDS)." heds.herts.ac.uk/, (acc. Oct 2000).

"Imaging & Document Solutions." Imaging Magazine, www.imagingmagazine.com, (acc. Oct 2000).

"International Federation of Library Association & Institutions (IFLA)." www.ifla.org/, (acc. Feb 2000).

"International Museum of Photography and Film, George Eastman House." George Eastman House, www.eastman.org, (acc. Oct 2000).

"JISC - Constitution and Funding Arrangements." Joint Information Systems Committee, www.jisc.ac.uk/jisc/const.html, (acc. Oct 2000), (created Sept 1997).

"National Archives and Record home page." www.nara.gov, (acc. Oct 2000).

"Photographic and Imaging Manufacturer's Association." PIMA, www.pima.net/standards/it10a.htm, (acc. Oct 2000), (created Aug 2000*).

"PRO home page." The Public Record Office, www.pro.gov.uk, (acc. Oct 2000).

"Products and Services Guide." Association for Information and Image Management (AIIM), www.aiimguide.com, (acc. Oct 2000).

"Programmes de numérisation du patrimoine culturel." Ministère Culture/Communication, www.culture.fr/culture/mrt/numerisation/fr/f_03.htm, (acc. Oct 2000), (created 1996).

"SCRAN home page." Scottish Cultural Resource Access Network, www.scran.ac.uk, (acc. Oct 2000), (created Sept 1996).

"Special Issue on Digital Libraries." *Communications of the Association for Computing Machinery (ACM)* 38, no. 4 (April 1995) (1995). www.acm.org/pubs/contents/journals/cacm/1995-38/, (acc. Oct 2000).

"The Image Permanence Institute." Rochester Institute of Technology, www.rit.edu/~661www1/, (acc. Oct 2000).

"Wilhelm Imaging Research Inc." Grinnell, Iowa, www.wilhelm-research.com, (acc. Oct 2000), (created June 2000*).

Collections, Content, and the Web. Washington, DC: Council on Library and Information Resources, 2000. www.clir.org/cpa/reports/pub88/contents.html, (acc. Oct 2000).

Arbuthnott, J. "Closing Remarks." In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 166-168.

Arms, C., J. Klavans, and D. J. Waters. *Enabling Access in Digital Libraries: A Report on a Workshop on Access Management.* Washington, DC: The Digital Library Federation, 1999.

Arms, W. Y. "Automated Digital Libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship?" *D-Lib Magazine* 6, no. 7/8, July/August (2000). www.dlib.org/dlib/july00/arms/07arms.html, (acc. Oct 2000).

Atkins, H. "The ISI® Web of Science® - Links and Electronic Journals. How links work today in the *Web of Scienc*e, and the challenges posed by electronic journals." *D-Lib Magazine* 5, no. 9, September (1999). www.dlib.org/dlib/september99/atkins/09atkins.html, (acc. Oct 2000).

Baker, N. "Deadline: The Author's Desperate Bid to Save America's Past." *The New Yorker* July 24 (2000): 42-61.

Barth, A., A. Endres, and A. de Kemp. *Digital Libraries in Computer Science.* Heigelberg, Germany: Springer-Verlag, 1997.

Bearman, D. "Trends in software for archives and museums." In *1994-95 Directory of Software for Archives and Museums*: Archives and Museum Informatics, 1994.

Bearman, D., (ed.) *Hypermedia and Interactivity in Museums: Proceedings of an International Conference 1991*. 1994.

Beckwith, B. R. "Issues Involved in Collections Management and Digitization of Images: a Review of the Literature." Drexel University, www.pages.drexel.edu/undergrad/brb24/rol.htm, (acc. Oct 2000), (created Feb 1999).

Besser, H., and J. Trant. *Introduction to Imaging: Issues in Constructing an Image Database.* Santa Monica, CA: The Getty Art History Information Program, 1995. www.getty.edu/gri/standard/introimages/index.html, (acc. Oct 2000).

Borgman, C. L. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World.* Cambridge, MA: MIT Press, 2000.

Borgman, C., (ed.) *Social Aspects of Digital Libraries. A Report on the UCLA-NSF Social Aspects of Digital Libraries Workshop (Los Angeles, Feb 15-17)*. 1996, Los Angeles, CA: Dept. of Library and Information Science, Univ. of California.

Bornstein, N. S. "Whose Net is it Anyway?" *Communications of the ACM* 41, no. 4, April 1998 (1998): 19-21.

Brilliant, R. "How an Art Historian Connects Art Objects and Information." *Library Trends* 37, no. 2 (1988): 120-129.

Brindley, L. "Developments in digitisation." In *Library Service Provision for Researchers: Proceedings of the Anderson report Seminar organised by the Library and Information Co-operation council (LINC) and the Standing Conference of National and University Libraries (SCONUL), Cranfield, 10 and 11 December 1996*: BRUTON, 1996.

Broadhurst, R. *MV- 39: The digitization of library material.* London: Library Information Technology Centre, 1993.

Bryant, P. *Making the most of our libraries: the report of two studies on the retrospective conversion of library catalogues in the United Kingdom and the need for a national strategy.*, *British Library Research & Innovation Centre Report 53*: British Library Research & Innovation Centre, 1997. www.ukoln.ac.uk/services/papers/bl/blri053/, (acc. Oct 2000).

Buckland, M. "The landscape of information science: The American Society for Information Science at 62."

Journal of the American Society of Information Science Special Issue JASIS at 50 (1999).

Burnard, L., M. Deegan, and H. Short, (eds.) *The Digital Demotic: A Selection of Papers from DRH97 (Digital Resources for the Humanities)*. 1997, London: Office for Humanities Communication.

Burnhill, P. “Enhancing Access, Ensuring Services: Data in the Digital Library.” In *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 59-81.

Carpenter, L. “Overview of the Digitisation Process.” In *Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September*, London: British Library, 1996. 2-5.

Chapman, A., N. Kingsley, and L. Dempsey. *Full Disclosure. Releasing the value of library and archive collections.*, *Library and Information Commission Report 10.* Bath: Library and Information Commission, 1999.

Chen, H. “Semantic Research for Digital Libraries.” *D-Lib Magazine* 5, no. 10, October (1999). www.dlib.org/dlib/october99/chen/10chen.html, (acc. Oct 2000).

Clews, J. “Digital Language Access: Scripts, Transliteration, and Computer Access.” *D-Lib Magazine* 3, no. 3, March (1997). www.dlib.org/dlib/march97/sesame/03clews.html, (acc. Oct 2000).

Commonwealth Higher Education Management Service. “Evaluation of the UK Pilot Site Licence Initiative - Phase II.” *HEFCE Ref 98/22,* 1998, National Information Services and Systems (NISS), www.niss.ac.uk/education/hefce/pub98/98_22.html, (acc. Oct 2000), (created May 1998).

Computer Science and Telecommunications Board. “Design and Evaluation: A Review of the State-of-the-Art.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/nrc/07nrc.html, (acc. Oct 2000).

Coppock, T., (ed.) *Making Information Available in Digital Format: Perspectives from Practitioners*. 1999, Edinburgh: The Stationary Office.

Corti, L., (ed.) *International Conference on Automatic Processing of Art History Data and Documents, September 24-27, 1984*. 1984, Pisa.

Cox, J., and M. Taleb. “Images on the Internet: Enhanced User Access.” *Database* 17, no. 4, August (1994): 18-22, 24-26.

Cummings, A. M., M. L. Witte, W. G. Bowea, L. O. Lazarus, and R. E. Ekman. *University Libraries and Scholarly Communication: Study Prepared for the Andrew W. Mellon Foundation.* Washington, DC: Association of Research Libraries, 1992.

Dale, R. L. “Introduction to Joint Conference Publications.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998.

Davis, B. “The Digital Museum and Computer-Generated Collections and Catalogs of Art.” *Aperture* 136, no. Summer (1994): 68-70.

Dearstyne, B. W. “What is the Use of Archives: A Challenge for the Profession.” *American Archivist* 50 (Winter) (1987): 76-87.

Deegan, M., C. Mullins, S. Ross, and S. Kenna, (eds.) *New Technologies for the Humanities*. 1996, London: Bowker-Saur.

Delaney, P., and G. Landow. *The Digital World.* Cambridge, MA 1993.

Department for Education and Employment. “Connecting the Learning Society: National Grid for Learning.” *The government’s consultation paper,* www.dfee.gov.uk/grid/consult/index.htm, (acc. Oct 2000), (created Oct 1997).

Department of National Heritage. *Reading the Future: a review of public libraries in England.*: Dept. of National Heritage, 1997.

Department of National Heritage. *Treasures in Trust: a review of museum policy.*: Dept. of National Heritage, 1996.

DeWitt, D. L., (ed.) *Going digital: strategies for access, preservation, and conversion of collections to a digital format*. 1998, New York: Haworth Press.

Elkington, N. “Joint RLG and NPO Conference on Guidelines for Digital Imaging.” *RLG DigiNews* 2, no. 5, October (1998). www.thames.rlg.org/preserv/diginews/diginews2-5.html#feature1, (acc. Oct 2000).

Elkington, N. E., (ed.) *Digital imaging technology for preservation*, *RLG Symposium on Digital Imaging Technology for Preservation*. 1994, Mountain View, CA: RLG.

Elkington, N. E., (ed.) *RLG Archives Microfilming Manual*. 1994, Mountain View, CA: The Research Libraries Group.

Elkington, N., (ed.) *RLG Preservation Microfilming Handbook*. 1992, Mountain View, CA: Research Libraries Group.

Elkington, N., and N. Beagrie. “After Event Report: Joint RLG and NPO Conference on Guidelines for Digital Imaging.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998.

Erway, R. L. “Digital Initiatives of the Research Libraries Group.” *D-Lib Magazine* 2, no. 12, December (1996). www.dlib.org/dlib/december96/rlg/12erway.html, (acc. Oct 2000).

Ester, M. *Digital Image Collections: Issues and Practices.* Washington, DC: Commission on Preservation and Access, 1996.

Feather, J. *Preservation and the Management of Library Collections.* London 1991.

Feeney, M., (ed.) *Digital culture: maximising the nations investment: a synthesis of JISC/NPO studies on the preservation of electronic materials*. 1999, London: British Library Board.

Fink, E. E. “The Getty Information Institute: A Retrospective.” *D-Lib Magazine* 5, no. 3, March (1999). www.dlib.org/dlib/march99/fink/03fink.html, (acc. Oct 2000).

Fleischhauer, C. “Digital Formats for Content Reproductions.” 1998, memory.loc.gov/ammem/formats.html, (acc. Oct 2000), (created July 1998).

Fox, E. A., and G. Marchionini, (guest eds.). “Toward a World Wide Digital Library.” *Communications of the ACM* 41, no. 4 (1998): 28-98.

Fox, E., R. Akscyn, R. Furuta, and J. Legget, (eds.). “Introduction to digital libraries.” *Communications of the ACM* 38, no. 4, April (1995): 22-28.

Fresko, M. “Current Technologies.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 5-8.

Fresko, M. “Rapportuer’s Overview.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 3.

Fresko, M. “Results of the Comparative Study of Digital Preservation Guidelines.” In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 46-51.

Fresko, M., (ed.) *Digitisation of Library Materials. Report of the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*. 1999: European Commission.

Fresko, M., (ed.). “Digitisation of Library Materials: Report on the Concentration Meeting & Workshop Luxembourg, 14.12.98.” Community Research and Development Information Service, ftp://ftp.cordis.lu/pub/ist/docs/digicult/digit.pdf, (acc. Oct 2000), (created Jan 1999).

Gassie, L. W., and N. Leroy. *Digitizing case studies: problems & solutions - on a grand scale!* Alexandria, VA: Audio Transcripts, Ltd, 2000. Audio cassette.

Gilliland-Swetland, A. J. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment.*: Council on Library and Information Resources, 2000.

Gould, C. C. *Information Needs in the Humanities: An Assessment.* Stanford, CA: The Research Libraries Group, 1988.

Gould, S., and R. Ebdon, (ed.). *IFLA/UNESCO Survey on Digitisation and Preservation.*, *International Preservation 2* 1999.

Graham, P. “Requirements for the Digital Research Library.” *College and Research Libraries* 56, no. 4 (1995): 331-339.

Graham, P. S. “New Roles for Special Collections on the Network.” *College and Research Libraries* 59, no. 3, May (1998): 232-239.

Graham, R. “Academic Image Exchange Formed.” *CLIR Issues* 10, no. July/August (1999). www.clir.org/pubs/issues/issues10.html#academic, (acc. Oct 2000).

Greenberger, M., (ed.) *Technologies for the 21st Century on Multimedia*. 1990, Santa Monica, CA: The Voyager Company.

Greenstein, D. “The Arts and Humanities Data Service: Three Years’ On.” *D-Lib Magazine* 4, no. 12, December (1998). www.dlib.org/dlib/december98/greenstein/12greenstein.html, (acc. Oct 2000).

Greenstein, D., and N. Beagrie. “Creating a Viable Scholarly Data Resource.” Arts and Humanities Data Service Web, www.ahds.ac.uk/deposit/viable.html, (acc. Oct 2000), (created Aug 1998*).

Greenstein, D., and N. Beagrie. “Draft Joint C&IT Policy.” 1999, Arts and Humanities Research Board, www.ahrb.ac.uk/strategy/index.htm, (acc. Oct 2000).

Griffin, S. M. “Digital Libraries Initiative - Phase 2: Fiscal Year 1999 Awards.” *D-Lib Magazine* 5, no. 7/8, July/August (1999). www.dlib.org/dlib/july99/07griffin.html, (acc. Oct 2000).

Griffin, S. M. “NSF/DARPA/NASA Digital Libraries Initiative: A Program Manager’s Perspective.” *D-Lib Magazine* 4, no. 7/8, July/August (1998). www.dlib.org/dlib/july98/07griffin.html, (acc. Oct 2000).

Hall, V. M. G. “Fair Use or Foul Play? The Digital Debate for Visual Resources Collections.” 1997, Johns Hopkins University, www.oberlin.edu/~art/vra/fairfoul.html, (acc. Oct 2000), (created Nov 1997).

Hare, G. “National Agency for Resource Discovery and the public library.” In *Towards a National Agency for Resource Discovery: Scoping Study, 1997*, edited by David Kay: British Library Research & Innovation Centre, 1997. 50-57.

Hare, G. “Public libraries in their political context.” In *The Internet, Networking and the Public Library*, edited by Sarah Ormes and Lorcan Dempsey: Library Association Publishing, published in association with UKOLN, 1997. 3-12.

Harter, S. P. “The Impact of Electronic Journals on Scholarly Communication: A Citation Analysis.” *The Public-Access Computer Systems Review* 7, no. 5 (1996). also at: http://info.lib.uh.edu/pr/v7/n5/hart7n5.html, (acc. Oct 2000).

Hawkins, B. L. “Creating the Library of the Future: Incrementalisation Won’t Get Us There!” In *New Serials: Proceedings of the American Serials Interest Group, Inc. (A special issue of the Serials Librarian)*, edited by Gail McMillan and Marilyn L Norstedt: The Haworth Press, Inc., 1994. 17-47.

Hearst, M. A. “Research in Support of Digital Libraries: at Xerox PARC Part I: The Changing Social Roles of Documents.” *D-Lib Magazine* 2, no. 5, May (1996). www.dlib.org/dlib/may96/05hearst.html, (acc. Oct 2000).

Hearst, M., G. Kopec, and D. Brotsky. “Research in Support of Digital Libraries at Xerox PARC Part II: Paper and Digital Documents.” *D-Lib Magazine* 2, no. 6, June (1996). www.dlib.org/dlib/june96/hearst/06hearst.html, (acc. Oct 2000).

Hemsley, J., and et al, (eds.) *Proceedings of the 1994 Electronic Imaging and the Visual Arts Conference, EVA ’94*. 1994, Fleet, UK: Brameur Ltd.

Hert, C. A., and G. Marchionini. “Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations.” 1987, University of North Carolina, http://ils.unc.edu/~march/blsreport/mainbls.html, (acc. Oct 2000), (created July 1997).

Hitchcock, S., L. Carr, W. Hall, S. Harris, S. Probets, D. Evans, and D. Brailsford. “Linking Electronic Journals: Lessons from the Open Journal project.” *D-Lib Magazine* 4, no. 12, December (1998). www.dlib.org/dlib/december98/12hitchcock.html, (acc. Oct 2000).

Hodge, G. M. *Systems of knowledge organization for digital libraries: beyond traditional authority files.* Washington, DC: Digital Library Federation, Council on Library and Information Resources, 2000.

Hodges, D., and C. D. Lunau. “The National Library of Canada’s Digital Library Initiatives.” *Library Hi Tech* 17, no. 2 (1999): 152-64.

Huttenlocher, D., and A. Moll. “On DigiPaper and the Dissemination of Electronic Documents.” *D-Lib Magazine* 6, no. 1, January (2000). www.dlib.org/dlib/january00/moll/01moll.html, (acc. Oct 2000).

IFLA. “Digital Libraries: Resources and Projects.” www.ifla.org/II/diglib.htm, (acc. Oct 2000), (created July 2000*).

Ihrig, S., and E. Ihrig. *Scanning the Professional Way.* Berkeley, CA: Osborne McGraw-Hill, 1995.

Information Technology in Humanities Scholarship: British Achievements, Prospects, and Barriers. London: The British Library Research and Development Department & The British Academy, 1993.

JISC. *Accessing our humanities collections: a guide to specialised collections for humanities researchers.* London: JISC, 1997.

Joint Funding Council’s Libraries Review. Report (The Follett Report). Bristol: HEFCE, 1993. www.ukoln.ac.uk/services/papers/follett/report/,

Kaser, R. T. “If Information Wants to Be Free . . . Then Who’s Going to Pay for It?” *D-Lib Magazine* 6, no. 5, May (2000). www.dlib.org/dlib/may00/kaser/05kaser.html, (acc. .

Kenney, A. R. “Americana Online: Products and Techniques.” 1999, University of Utrecht, www.cs.uu.nl/events/dech1999/kenney/index.htm, (acc. Oct 2000), (created Dec 1999).

Kenney, A. R. “Guidelines vs. Guidance for Digital Imaging: The Opportunity Before Us.” In *Guidelines for Digital Imaging: papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998*, London: The National Preservation Office, 1998. 1-10.

Kenney, A. R., and O. Y. Rieger, (eds.) *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. 2000, Mountain View, CA: Cornell University Press.

Kenney, A. R., and O. Y. Rieger. *Managing Digital Imaging Projects: An RLG Workshop.*: RLG, 1998.

Kenney, A. R., and S. Chapman, (eds.) *Digital Imaging for Libraries and Archives*. 1996, Ithaca, NY: Cornell University Library Press.

Klavans, J. L. “New Center at Columbia University for Digital Library Research: Fostering Interdisciplinary Research and Bridging Cultural Clashes.” *D-Lib Magazine* 2, no. 3, March (1996). www.dlib.org/dlib/march96/klavans/03klavans.html, (acc. Oct 2000).

Knoll, A. “For Whom and How We Digitize.” *Web World, Memory of the World, Memoriae Mundi Series, Bohemica,* 1998, UNESCO, www.unesco.org/webworld/mdm/czech_digitization/doc/users.htm#ForWhom%20To, (acc. Oct 2000).

Lawrence, G. W., W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney. *Risk Management of Digital Information: A File Format Investigation.* Washington, DC: Council on Library and Information Resources, 2000. www.clir.org/pubs/reports/pub93/contents.html, (acc. Oct 2000).

LC21: A Digital Strategy for the Library of Congress. Washington, DC: National Academy Press, 2000. http://www.nap.edu/books/0309071445/html/, (acc. Oct 2000).

Leggat, R. “A History of Photography.” KBnet, www.kbnet.co.uk/rleggat/photo/index.html, (acc. Oct 2000), (created Dec 1995).

Lesk, M. “Perspectives on DLI-2 - Growing the Field.” *D-Lib Magazine* 5, no. 7/8, July/August (1999). www.dlib.org/dlib/july99/07lesk.html, (acc. Oct 2000).

Lesk, M. “Why digital libraries? (Follett Lecture Series).” 1995, UK Office for Library and Information Networking, www.ukoln.ac.uk/services/papers/follett/lesk/paper.html, (acc. Oct 2000), (created June 1995).

Lesk, M. *Practical Digital Libraries: Books, Bytes, and Bucks.* San Francisco: Morgan Kaufmann Publishers, Inc., 1997.

Levy, D. M. “What Do You See and What Do You Get? Document Identity and Electronic Media.” In *Screening Words: User Interfaces for Text; Proceedings of the Eighth Annual Conference of the UV Centre for the New OED and Text Research*, 1992. 109-117.

Lindenbeck, J. “Digitale Bildearbeitung.” *Archaeologisches Nachrichtenblatt* 2, no. 1 (1997): 3-21.

Lynch, C., and L. Lunin. “Perspectives: Advanced Applications of Imaging.” *Special issue of Journal of the American Society for Information Science* September (1991).

Matheson, A. “Conclusions and recommendations.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 91-93.

McElhone, J. “Electronic Imaging: Information Gathered at Two Events.” *Spectra* 25, no. 1 (1997): 48-56.

Menne-Haritz, A., and N. Brübach. “The Intrinsic Value of Archive and Library Material.” Archivschule Marburg, Digitaler Texte Nr. 5, www.uni-marburg.de/archivschule/intrinsengl.html, (acc. Oct 2000), (created Aug 1999).

Michalko, J. *Equilibrium and Opportunism. Information Strategies and the New Environment.*: New Library World, *forthcoming*.

Michelson, A., and J. Rothenberg. “Scholarly Communication and Information Technology: Exploring the Impact of Changes in the Research Process on Archives.” *American Archivist* 55, no. 2 (1992).

Mitchell, W. J. *The Reconfigured Eye: Visual Truth in the Past-Photographic Era.* Boston: Massachusetts Institute of Technology, 1992.

Mulder, H., and F. Wiering, (eds.). “Conference Papers.” *Digitisation of European Cultural Heritage: Products - Principles - Techniques (21-23 October Utrecht, The Netherlands),* University of Utrecht, www.cs.uu.nl/events/dech1999/dech.htm, (acc. Oct 2000), (created Feb 2000*).

Okerson, A., and D. Mogge, (eds.) *Gateways, Gatekeepers, and Roles in the Information Omniverse, Proceedings from the Third Symposium*. 1993, Washington, DC: Association of Research Libraries - Association of American University Presses.

Okerson, A., and J. J. O’Donnell. *Scholarly journals at the crossroads: a subversive proposal for electronic publishing.* Washington, DC: Office of Scientific & Academic Pub., Association of Research Libraries, 1995.

Parry, D. *Virtually New. Creating the Digital Collection. A Review of Digitisation Projects in Local Authority Libraries and Archives.* London: Library and Information Commission, 1998.

Peek, R., (ed.). “Special Topic Issue: Electronic Publishing.” *Journal of the American Society for Information Science* 47, no. 9 (1996).

Preservation and Digitisation: principles, practice and policies. Papers given at the National Preservation Office 1996 Annual Conference, University of York, 3-5 September. London: British Library, 1996.

Rice, S. V., J. Kanai, and T. A. Nartaker. “The Third Annual Test of OCR Accuracy.” In *ISRI Technical Report no. 94-03 April 1994 (presented at SDAIR ’94, the Third Annual Symposium on Document Analysis and Information Retrieval)*, 1994. 11-28.

Schwartz, G. “Digital Imagery and User-defined Art.” *Art Bulletin* 79, no. 2 (1997): 206-208.

Smith, A. *Why Digitize?* Washington, DC: CLIR, 1999. www.clir.org/pubs/reports/pub80-smith/pub80.html, (acc. Oct 2000).

Strunck, K., and J. Thorhauge. “Three stories from the future.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 69-76.

Sundt, C. L., (ed.),. “Issues in Electronic Imaging.” *Special issue of Visual Resources* 10, no. 1 (1994).

The Southeastern Library Network Inc. “Selected Resources on Digital Imaging for Libraries and Archives.” SOLINET, www.solinet.net/presvtn/leaf/imaging.htm, (acc. Oct 2000), (created Jan 2000*).

Tupek, A. R., and C. S. Dippo. “Quantitative Literacy: New Website for Federal Statistics Provides Research Opportunities.” *D-Lib Magazine* 3, no. 12, December (1997). www.dlib.org/dlib/december97/stats/12tupek.html, (acc. Oct 2000).

Wiering, F., and H. Mulder, (eds.) *Digitisation of European Cultural Heritage*. 2000, Utrecht: Netherlands Utrecht University.

previous content next

valid xhtml 1.1
abp~03/02

**The NINCH Guide to Good Practice
in the Digital Representation and Management
of Cultural Heritage Materials**

Background Bibliography

13. Standards

"AMICO Data Specification." AMICO, www.amico.net/docs/dataspec.html, (acc. Oct 2000).

"An Annotated Bibliography of Interoperability Literature." Stanford University, www-diglib.stanford.edu/diglib/pub/interopbib.html, (acc. Oct 2000).

"California Digital Library: Digital Collection Image Standards." 1999, California Digital Library, www.ucop.edu/irc/cdl/tasw/Current/Imaging.Stds-090199/Imaging.Stds-090199.doc, (acc. Oct 2000), (created Sept 1999).

"Color Management." Dunway Products, www.dunway.com/secties/Information/ColorManagement/ColorManagement.html, (acc. Jan 2000).

"Color Scanners: How We Tested." *PC Magazine Online*, no. March 4 (1997). www.zdnet.com/pcmag/features/scanners/hwt.htm, (acc. Oct 2000).

"Colour Characteristics of Digital Still Cameras (DSCs) Using Colour Targets and Spectral Illumination." *Program of Work,* PIMA/IT10: Electronic Still Picture Imaging, www.pima.net/standards/it10/IT10_POW.htm, (acc. Oct 2000), (created Sept 2000*).

"Consortium for the Computer Interchange of Museum Information." www.cimi.org, (acc. Oct 2000), (created Aug 2000).

"Data Attributes." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* lcweb.loc.gov:8081/ndlint/repository/attribs.html, (acc. Oct 2000), (created May 1997).

"Data Definitions." *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* lcweb.loc.gov:8081/ndlint/repository/attdefs.html, (acc. Oct 2000), (created May 1997).

"Data Standards." Visual Resources Association, www.oberlin.edu/~art/vra/dsc.html, (acc. Oct 2000), (created June 2000).

"DC Date Qualifiers: DC (Final) Working Draft - 14 December 1999." 1999, Dublin Core Metadata Initiative, www.mailbase.ac.uk/lists/dc-usage/files/date-final.html, (acc. Oct 2000), (created December 1999).

"Developments in museum and cultural heritage information standard." International Committee for Documentation of the International Council of Museums (ICOM-CIDOC)/Getty Information Institute, www.cidoc.icom.org/stand1.htm, (acc. Oct 2000), (created July 1996).

"Digital Camera Reviews." Digital Photography Reviews, www.inconference.com/digicam/camera.html, (acc. Oct 2000).

"Digital Image Basics: Some Explanations and Descriptions of Terms." *Digital Images in Multimedia Presentations,* Technical Advisory Service for Images (TASI), www.ilrt.bris.ac.uk/mru/dimp/section3/basics_prt.html, (acc. Oct 2000).

“Digital imaging and optical storage technology helps open up railway archives at National Railway Museum.” *Electronic Library* 11, no. 2 (1993): 108-109.

“Digital Libraries: Metadata Resources.” IFLANET, www.ifla.org/II/metadata.htm, (acc. Oct 2000), (created Oct 2000*).

“Digital Page Imaging and SGML: An Introduction to the Electronic Binding DTD (Ebind).” Berkeley Digital Library SunSITE, sunsite.berkeley.edu/Ebind/index.html, (acc. Oct 2000), (created June 1996).

“Dublin Core Metadata Initiative home page.” purl.oclc.org/dc, (acc. Oct 2000), (created Oct 1998).

“Encoded Archival Description Official Web Site.” Library of Congress, Network Development & MARC Standards Office, lcweb.loc.gov/ead, (acc. Oct 2000), (created Sept 2000*).

“Guide to Best Practice: Dublin Core.” 1999, Consortium for Computer Interchange of Museum Information, www.cimi.org/documents/meta_bestprac_final_0899.pdf, (acc. Jan 2000), (created Aug 1999).

“Guidelines and Standards: Getting Started.” Colorado Digitization Project, coloradodigital.coalliance.org/standard.html, (acc. Jan 2000), (created 1999).

“Image File Formats List.” Center for Innovative Computer Applications, http://cica.indiana.edu/graphics/image.formats.html, (acc. Jan 2000), (created Jan 1994).

“ISO Archiving Standards - Reference Model Papers.” ssdoo.gsfc.nasa.gov/nost/isoas/, (acc. Oct 2000), (created July 2000*).

“KODAK PHOTO CD Technology/Transferring Images.” *Kodak Service & Support Frequently Asked Questions,* www.kodak.com/cluster/global/en/service/faqs/faq1001c.shtml, (acc. Oct 2000), (created Nov 1998*).

“Library of Congress Maintenance Agency Page for International Standard Z39.50.” Library of Congress, lcweb.loc.gov/z3950/agency, (acc. Oct 2000), (created Aug 2000*).

“MARC 21 Concise Bibliographic: Control Field 007.” Library of Congress, lcweb.loc.gov/marc/bibliographic/ecbd007s.html#mrcb007c, (acc. Oct 2000), (created Jan 2000*).

“MARC Standards.” Library of Congress, Network Development & MARC Standards Office, lcweb.loc.gov/marc, (acc. Oct 2000), (created June 2000*).

“Monitor Adjustment Target.” National Archives and Records Administration, NARA Archival Information Locator (NAIL), www.nara.gov/nara/target.html, (acc. Jan 2000), (created March 2000*).

“Munsell Color Vision Tests.” GretagMacbeth, www.munsell.com/munvis.htm, (acc. Jan 2000).

“NISO/CLIR/RLG Technical Metadata Elements for Images Workshop.” 1999, NISO (National Information Standards Organization), www.niso.org/image.html, (acc. Oct 2000), (created April 1999).

“Resource Description Framework (RDF).” Word Wide Web Consortium, www.w3.org/RDF/, (acc. Oct 2000).

“Sample Structural Metadata for the Coolidge Image-Only Documents.” *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/timag-samp.html, (acc. Oct 2000), (created March 1997).

“Sample Structural Metadata for the Coolidge Photos.” *Structural Metadata Dictionary for Library of Congress Repository Digital Objects,* Library of Congress, lcweb.loc.gov:8081/ndlint/repository/photo-samp.html, (acc. Oct 2000), (created March 1997).

“Setting the Resolution.” *Photography Collection,* Denver Public Library, Western History/Genealogy Department, gowest.coalliance.org/calib.htm, (acc. Oct 2000).

“Technical Recommendation for Digital Imaging Projects.” *Image Quality Working Group of ArchivesCom, a joint Libraries/AcIS Committee,* 1997, University of Columbia, www.columbia.edu/acis/dl/imagespec.html, (acc. Oct 2000), (created April 1997*).

“The Conversion Process.” Göttinger Digitalisierungs-Zentrum, www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html#conversion, (acc. Oct 2000), (created June 1999).

"The Data Documentation Initiative: A Project to Develop an XML Document Type Definition for Data Documentation." ICPR, www.icpsr.umich.edu/DDI/code-book.html, (acc. Oct 2000), (created March 2000).

Bicknese, D. "Measuring the Accuracy of the OCR in the Making of America." *(a report prepared in fulfillment of directed field experience requirements, University of Michigan, School of Information, winter 1998),* University of Michigan, moa.umdl.umich.edu/moaocr.html, (acc. Oct 2000), (created 1998).

Acland, G., K. Cumming, and S. McKemmish. "The End of the Beginning: The SPIRT Recordkeeping Metadata Project." *Australian Society of Archivists 1999 Conference,* 1999, www.archivists.org.au/events/conf99/spirt.html, (acc. Oct 2000).

American Society for Testing Materials. *ASTM Standards on Color and Appearance Measurement.* 5th ed. West Conshohocken, PA: ASTM, 1996.

ANSI/AIIM TR26-1993, Resolution as it Relates to Photographic and Electronic Imaging. Silver Spring, MD: Association for Information and Image Management, 1993.

Baca, M., "Introduction to Metadata. Pathways to Digital Information." Getty Institute, www.getty.edu/gri/standard/intrometadata/, (acc. Oct 2000).

Baca, M., and Patricia Harping (guest eds.). "Art Information Task Force: Categories for the Descriptions of Works of Art." *Visual Resources* 11, no. 3-4 (1996): 41-429.

Bach, J. R., C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu. "The Virage Image Search Engine: An Open Framework for Image Management." In *Storage and Retrieval for Image and Video Databases IV, SPIE Proceedings 2670*, Bellingham, WA: SPIE, 1996. 76-87.

Baker, T. "Languages for Dublin Core." *D-Lib Magazine* 4, no. 12, December (1998). www.dlib.org/dlib/december98/12baker.html, (acc. Oct 2000).

Beagrie, N., and D. Greenstein. "A Strategic Policy Framework for Creating and Preserving Digital Collections." 1998, Arts and Humanities Data Service, www.ahds.ac.uk/manage/framework.htm, (acc. Jan 2000).

Beagrie, N., and D. Greenstein. "Managing Digital Collections: AHDS Policies, Standards and Practices Version 1." 1998, Arts and Humanities Data Service, www.ahds.ac.uk/public/srg.html, (acc. Oct 2000), (created Dec 1998).

Bearman, D. "NISO/CLIR/RLG Technical Metadata Elements for Images Workshop, April 18-19, 1999." NISO (National Information Standards Organization), www.niso.org/imagerpt.html, (acc. Oct 2000), (created April 1999).

Bearman, D. A., and J. Perkins, (eds.) *Standards Framework for the Computer Interchange of Museum Information*. 1st ed, *Museum Computer Network*. 1993.

Bearman, D., and J. Trant. "Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process." *D-Lib Magazine* 4, no. 6, June (1998). www.dlib.org/dlib/june98/06bearman.html, (acc. Oct 2000).

Bearman, D., and K. Sochats. "Metadata Requirements for Evidence." University of Pittsburgh, www.lis.pitt/edu/~nhprc/BACartic.html, (acc. Oct 2000).

Bearman, D., G. Rust, S. Weibel, E. Miller, and J. Trant. "A Common Model to Support

. Progress Report on Reconciling Metadata Requirements from the Dublin Core and INDECS/DOI Communities." *D-Lib Magazine* 5, no. 1, January (1999). www.dlib.org/dlib/january99/bearman/01bearman.html, (acc. Oct 2000).

Brown, C. W., and B. J. Shephard. *Graphics File Formats, Reference and Guide.* Upper Saddle River, NJ: Prentice Hall PTR, 1995.

Buckland, M., A. Chen, H.-M. Chen, Y. Kim, B. Lam, R. Larson, B. Norgard, and J. Purat. "Mapping Entry Vocabulary to Unfamiliar Metadata." *D-Lib Magazine* 5, no. 1, January (1999). www.dlib.org/dlib/january99/buckland/01buckland.html, (acc. Oct 2000).

Bureau of Canadian Archivists. *Planning Committee on Descriptive Standards. Subject Indexing for Archives.*

The Report of the Subject Indexing Working Group. Ottawa: Bureau of Canadian Archivists, 1992.

Consortium for the Computer Interchange of Museum Information CIMI Z39.50 Working Group. “The CIMI Profile: Z39.50 Application Profile for Cultural Heritage Information.” 1988, www.cimi.org, (acc. Aug 2000).

Dempsey, L., and S. L. Weibel. “The Warwick Metadata Workshop: a framework for the deployment of resource description.” *D-Lib Magazine* 2, no. 7/8, July/August (1996). www.dlib.org/dlib/july96/07weibel.html, (acc. Oct 2000).

Engel, V. S. “The Browser Safe Palette (BSP).” www.the-light.com/netcol.html, (acc. Oct 2000), (created 1996).

Frey, F. S., and J. M. Reilly. *Digital Imaging for Photographic Collections: Foundations for Technical Standards.* Rochester, NY: Image Permanence Institute, Rochester Institute of Technology, 1999. www.rit.edu/~661www1/sub_pages/page17.pdf, (acc. Oct 2000).

Gann, R. G. *Desktop Scanners: Image Quality Evaluation.* Upper Saddle River, NJ: Prentice-Hall, 1999.

Gann, R., J. Goode, and R. Wadas. *Reviewing and Testing Desktop Scanners.* Second ed. Greeley, CO: Hewlett Packard Company, 1995.

Gillings, M., A. Wise, and (eds). “GIS Guide to Good Practice.” 1998, Arts and Humanities Data Service, London, ads.ahds.ac.uk/project/goodguides/gis/index.html, (acc. Oct 2000).

Gillings, M., and A. Wise, (eds.) *GIS Guide to Good Practice*. 1999, Oxford: Oxbow Books for the Arts and Humanities Data Service.

Gillman, P. *National Name Authority Files: a report to the National Council on Archives.*, *BL Research & Innovation Report 91, 1998*: Historical Manuscripts Commission, 1998. www.hmc.gov.uk/nca/nnaftit.htm, (acc. Oct 2000).

Gordon, C. “An Introduction to ICONCLASS.” In *Terminology for Museums: Proceedings of an International Conference Held in Cambridge, England, 21-24 September 1988*, edited by D Andrew Roberts, Cambridge, UK: Museum Documentation Association with the Assistance of the Getty Grant Program, 1990. 233-244.

Gorman, M., (ed.) *Anglo American Catologuing Rules 2*. second ed. 1989, Chicago: Library Association.

Guidelines for Digital Imaging: Papers given at the Joint National Preservation Office and Research Libraries Group Preservation Conversation in Warwick 28th - 30th September 1998. London: The National Preservation Office, 1998. www.rlg.ac.uk/preserv/joint/, (acc. Oct 2000).

Hakala, J. “Metadata and identifiers for e-journals.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 53-68.

Helmuth, N. “Digital Photography.” www.digital-photography.org/default.html, (acc. Oct 2000), (created Dec 1999*).

Hillmann, D. I. “Choices: MARC or Dublin Core?” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 89-90.

Hofman, P., E. Worsfold, D. Hiom, M. Day, and A. Oehler. “Specification for Resource Description Methods: 2, Selection Criteria for Quality Controlled Information Gateways.” *DESIRE: Development of a European Service for Information on Research and Education, Deliverable 3.2(2),* 1997, www.ukoln.ac.uk/metadata/desire/quality, (acc. Oct 2000), (created May 1997).

Holm, J. “Survey of Developing Electronic Photography Standards.” *Critical Reviews of Optical Science and Technology, SPIE* CR61 (1996): 120-152.

International Council on Archives. *ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families.*: ICA, 1995.

International Council on Archives. *ISAD(G): General International Standard Archival Description. Adopted by the Ad Hoc Commission on Descriptive Standards.* Stockholm, Sweden 1993.

International Organization for Standardization. *Graphic Technology - Prepress Digital Data Exchange - Tag Image File Format for Image Technology (TIFF/IT).*, *Technical Committee 130, ISO/FDIS 12639.* Geneva: International Organization for Standardization, 1998.

ISAAR (CPF). *International Standard Archival Authority Record for Corporate Bodies, Persons, and Families. Final ICA approved version. Prepared by the AD HOC Commission on Descriptive Standards.* Paris 1995.

ISO 12233/FDIS, ISO/TC42, Photography - Electronic Still Picture Cameras - Resolution Measurements. 1999.

ISO 14524/FDIS, ISO/TC42, Photography - Electronic Still Picture Cameras - Methods for Measuring Opto-Electronic Conversion Functions (OECFs). 1999.

ISO 15739/CD, ISO/TC42, Photography - Electronic Still Picture Cameras - Noise Measurements. 1999.

ISO 16067/WD, ISO/TC42, Electronic scanners for photographic images - Spatial resolution measurements: Part 1, Scanners for reflective media. 1999.

ISO 17321/WD, ISOTC42, Graphic Technology and Photography - Colour Characterisation of Electronic Still Cameras (DSCs) Using Colour Targets and Spectral Illumination. 1999.

ISO/DIS 12234-2, Photography - Electronic Still Pictures Cameras - Removable Memory - Part 2: Image Data Format - TIFF/EP. 1998.

ISO/FDIS 12233 Photography, Electronic Still Picture Cameras, Resolution Measurements.

ISO/IEC. “IS 12087-5, Information Technology - Computer Graphics and Image Processing - Image Processing Interchange (IPI) - Functional Specification - Part 5: Basic Image Interchange Format (BIIF).” 1998, http://164.214.2.51/ntb/related/, (acc. Jan 2000), (created Dec 1998).

Joint Funding Councils Non-Formula Funding of Specialised Research Collections in the Humanities. “Z39.50 for Archival Applications.” *Scoping and Draft Specification Document: Fretwell-Downing Informatics,* 1996, www.kcl.ac.uk/projects/srch/reports/z3950.htm, (acc. Aug 2000).

Lagoze, C., and S. Payette. “Metadata: Principles, Practices, and Challenges.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 84-100.

Lassila, O., and R. Swick, (eds.) *World Wide Web Consortium, 1998, Resource Description Framework (RDF) Model and Syntax Specification. WC3 Working Draft*. 1998.

Library of Congress. “The Z39.50 Profile for Access to Digital Collections.” 1997, lcweb.loc.gov/z3950/agency/profiles/collections.html, (acc. Oct 2000), (created May 1996).

Lossau, N., and F. Klaproth. “TIFF Header: A Reference Stamp for Image Files.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 97-98.

MARC, The Machine Readable Catalogue Record, Specified as ISO 2709.

McCamy, C. S., H. Marcus, and J. G. Davidson. “A Color Rendering Chart.” *Journal of Applied Photographic Engineering* 2 (1976): 95-99.

Mehrabi, H., and H. Laursen. “Standards for images and full text.” In *Digitising Journals. Conference on future strategies for European libraries proceedings*, edited by Vibeke Cranfield, Copenhagan: Danish National Library Authority/Denmark’s Research Library, 2000. 43-46.

Miller, E. “An Introduction to the Resource Description Framework.” *D-Lib Magazine*, no. May (1998). www.dlib.org/dlib/may98/miller/05miller.html, (acc. Oct 2000).

Miller, P. “Metadata for the Masses.” *Ariadne* 5. www.ariadne.ac.uk/issue5/metadata-masses/intro.html, (acc. Oct 2000).

Morrison, I. “SCRAN Basic Record Standard.” 1997, Scottish Cultural Resource Access Network (SCRAN), Edinburgh, www.scran.ac.uk/articles/scranbds.htm, (acc. Oct 2000).

Morrison, I. “SCRAN Data Capture Standards.” 1997, SCRAN (Scottish Cultural Resource Access Network),

Edinburgh, www.scran.ac.uk/articles/scrancds.htm, (acc. Oct 2000).

MPEG-7 Requirements Group. *MPEG-7 Requirements Document V.9.*, *July 1999 ISO/IEC SC29/WG11 N2859.* Vancouver 1999.

MPEG-7 Requirements Group. *MPEG-7: Context, Objectives and Technical Roadmap, V.12.*, *July 1999 ISO/IEC SC29/WG11 N2861.* Vancouver 1999.

MPEG-7 Video Group. *MPEG-7 Visual Part of the eXperimentation Model Version 2.0.*, *ISO/IEC SC29/WG11 N2822.* Vancouver 1999.

Museum Documentation Association. *MDA Archaeological Objects Thesaurus.* Cambridge: Museum Documentation Association, English Heritage, and the Royal Commission for the Historical Monuments of England, 1997.

Museum Documentation Association. *SPECTRUM: The UK Museum Documentation Standard.* 2nd ed. Cambridge: Museum Documentation Association, 1997.

NMEA. *NMEA 0183 Standard.* New Bern, NC: National Maritime Electronics, 1995.

Oile, T. "Impact of standardization work on the future of information technology." In *IFIP World Conference on IT Tools*, edited by N Terashima and E Altman, New York: Chapman & Hall, 1996. 97-105.

Paskin, N. "DOI: Current Status and Outlook." *D-Lib Magazine* 5, no. 5, May (1999). www.dlib.org/dlib/may99/05paskin.html, (acc. Oct 2000).

Paskin, N., and G. Rust. "The Digital Object Identifier Initiative: Metadata Implications." *DOI Discussion Paper 2, version 3,* 1999, dx.doi.org/10.100/104, (acc. Oct 2000), (created Feb 1999).

Perkins, J. "The CIMI standard for the interchange of multimedia information." In *Museums and Interactive Multimedia: Proceedings of an International Conference held in Cambridge, England, 20-24 September 1993 (6th International Conference on Hypermedia and Interactivity in Museums)*, edited by Diane Lees and Museum Documentation Association, Cambridge: MDA, 1993. 244-254.

Perlman, E., and I. Kallen. "Common Internet File Formats." 1995, www.matisse.net/files/formats.html, (acc. Oct 2000), (created Dec 1995*).

Puglia, S. "Fractal and Wavelet Compression." *RLG DigiNews* 2, no. 3 (1998). www.rlg.org/preserv/diginews/diginews23.html#technical2, (acc. Oct 2000).

Rabbani, M. "Image Compression Fundamentals." In *The Compression Experience: Proceedings of The Rochester Section of the SMPTE (Society of Motion Picture & Television Engineers) Tutorial*, Rochester, NY, 1995. 7-24.

Resolution as it Relates to Photographic and Electronic Imaging. ANSI/AIIM TR26-1993, *Technical Report.* Silver Spring, MD: Association for Information and Image Management, 1993.

Ross, S. "From conventional photographs to Digital Resources." *Archaeological Computing Newsletter* 35 (1993): 14-21.

Rust, G., and M. Bide. "The <indecs> Metadata Model." *version no. 3,* 1999, indecs, www.indecs.org/results/archive.htm, (acc. Oct 2000), (created July 1999).

Standards in the Museum Care of Photographic Collections.: Museums and Galleries Commission, 1996.

Steele, M., J. Murdoch, and M. Arnott. "The Emergence of Photo-CD as a Preferred Method for Image Capture and Storage." *Information Services and Use* 16 (1996): 9.

Stokes, M., M. Anderson, S. Chandrasekar, and R. Motta. "A Standard Default Color Space for the Internet - sRGB." *Version 1.10,* 1996, www.w3.org/Graphics/Color/sRGB, (acc. Oct 2000), (created Nov 1996).

Tanner, S. "Standards." In *Digitisation of Library Materials. Report on the Concertation Meeting and Workshop, held in Luxembourg, 14 December 1998*, edited by Marc Fresko: European Commission, 1999. 9-10.

Turnbull, G., and A. Buchanan. "SCRAN Project Guidelines." 1997, Scottish Cultural Resource Access Network, Edinburgh, www.scran.ac.uk/articles/projpaper.htm, (acc. Oct 2000), (created May 1997*).

Vitale, T. “Light Levels Used in Modern Flatbed Scanners.” *RLG DigiNews* 2, no. 5, 15 October (1998). www.rlg.org/preserv/diginews/diginews2-5.html#technical, (acc. Oct 2000).

Wenman, L. “The Browser-Safe Color Palette.” www.lynda.com/hex.html, (acc. Oct 2000).

Williams, D. “An Overview of Image-Quality Metrics.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 72-73.

Williams, D. “What is an MTF...and Why Should You Care?” *RLG DigiNews* 2, no. 1, February 15 (1998). www.rlg.org/preserv/diginews/diginews21.html#technical, (acc. Oct 2000).

World Wide Web Consortium. “Synchronized Multimedia Integration Language.” www.w3.org/AudioVideo/, (acc. Oct 2000).

Wu, Y. “Raster, Vector, and Automated Raster-to-Vector Conversion.” In *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, edited by Anne R Kenney and Oya Y Rieger, Mountain View, CA: Cornell University Library, 2000. 117-118.

X-Rite®. *The Color Guide and Glossary: Communication, Measurement, and Control for Digital Imaging and Graphic Arts.* 1998. www.x-rite.com/Documents/Mktg/L11-029.pdf, (acc. Oct 2000).

previous content

valid xhtml 1.1

abp~03/02

2 The Bancroft Library

HATII interviewed Merrilee Proffitt, Director of Digital Archives Development at The Bancroft Library, University of California, on January 25 2001. A result of the California Heritage Digital Image Access Project, images are selected from the Manuscript and Pictorial Collections of The Bancroft Library, to then be digitized and contained in the Online Archive of California. The OAC began as a pilot project, with the aim of developing a UC-wide database of archival finding aid data. The Bancroft began the process of digitizing its thirty-nine volumes of historical records, for the purposes of enhancing research, teaching and learning opportunities, while also opening up the resources for greater public access.

2.1 Organizational Digitization Program and Policy

The Bancroft Library, like other collections in the Online Archive of California, has access to its rich digital material through the OAC central server via the finding aids produced in Bancroft.

The collections at Bancroft contain material in special collections such as manuscripts and rare books. Among its components are the Bancroft Collection of Western Americana and Latin Americana, the Rare Book Collection, the History of Science and Technology Collection, the University Archives, the Free Speech Movement Project, the Bancroft Library Pictorial Collection, the Mark Twain Papers and Project, and the Regional Oral History Office. There is no formal collection survey for digitization, which tends to take place in an opportunistic way, i.e. in response to funding opportunities, or faculty alliances on campus, which the Bancroft tries to cultivate to produce digital resources that are directly related to UC Berkeley research. A curators' management group meets quarterly and discusses various issues, including which possible digital projects could be pursued based on the opportunities available. The catalog is examined for projects, which may appear to be feasible on paper, but could be difficult in practice, i.e. those involving fragile objects or copyright issues. The research is often focused on regional and local thematic bases.

They will often incorporate funding in any grant for collection surveys specifically to ensure that all relevant material is identified. The finding aids associated with the collection, whether paper or electronic, are essential to any project at all stages.

There is a thematic approach to the selection of material for digitization, based on the research projects identified. The material will be related to the subject area. However, the nature of special collection material is such that it can often pose physical problems in digitizing, so the selection process can be very subjective. While Bancroft does not have formal selection criteria for digitization, it recognizes that such a document would be most useful – although not a restrictive document, rather one that allowed flexibility with guidelines. The management group functions as a "living" document on selection criteria.

The main influences are research significance and teaching and learning potential. They are not driven by labor or infrastructure cost reduction, or by space rationalization. With selection a response to research, the criteria change over time and from project to project.

As with sub-projects in the OAC, the Bancroft has co-operated with various projects across the cultural heritage field, as well as with corporations such as Biotech companies. Collaboration is considered to be worthwhile if partners are chosen carefully and roles and commitments are set out early, while having reasonable expectations of members. Face to face meetings are vital and should happen soon and often. It is essential to remember that the goal is to produce and to collaborate.

The Bancroft started digitization in 1993 when it converted the card catalog to MARC on EAD. The principal purposes behind the whole digitization project are research, teaching and learning, and public access, as would be expected from such an institution. Materials digitized include handwritten documents, typescript documents, original artworks and photographs, however, the nature of these materials varies, reflecting the diverse nature of the collection as a whole. They rarely digitize all of one sub-collection, i.e. one leaf from each manuscript in a collection. They did fully digitize their papyrus collection, but this was not typical.

The finding aids must be able to be used by OAC for access on the OAC central server, so the Bancroft uses standards where possible, including CIMI, TEI, JPEG, SGML and XML. Other standards for describing content are MARC, EAD, TEI Header and CDWA. They have an “adopt and adapt” approach, using standards where possible and adapting where practically necessary; for example, the papyrus project had to create a thesaurus of ancient Egypt place names. Interestingly, they are not enthusiastic about Dublin Core, but this could be a librarian’s perspective, where Dublin Core is not as flexible as MARC and or EAD.

2.2 Project Management and Planning

Project management varies from project to project, but tends to be in-house and responds to the needs of individual projects. The management group is key to the digitization project management, meeting twice a week to discuss all Library activities and quarterly reports on the digital issues. Within projects, email is used to track progress and goals. The Bancroft is keen to establish more formal project management procedures for digital projects and recognizes the need for training on project management. At the moment, such training is geared towards business rather than cultural and heritage activities.

They do not carry out pilot studies or feasibility studies prior to starting projects but often base benchmarks on previous projects. Text conversion for EAD and TEI is out-sourced, as is microfilm, while imaging of the collections material is carried out in-house. There is an excellent imaging unit in Bancroft, which was unfortunately unavailable to this visit, due to illness. In this unit, run by Dan Johnston, fragile and delicate materials can be imaged with great care. Flatbed scanners handle oversized materials and there is a Phase One - Power Phase camera. This unit uses Mac workstations and has its own intranet for managing digital objects.

2.3 Human Resources and Training

Staffing on each project varies, with archivists and curators involved at different levels. For example, more curators were involved on the papyrus project due to the nature of the material, and students are often employed where there is a large amount of material to process. In the Etext unit there are three programmers and three text encoding specialists. Training on digitization is provided in-house by Dan Johnston and his team. Staff members are also offered training on using the databases for cataloging.

Bancroft feels that there is a huge need for the community to offer more specific training on all issues associated with digitization, in short intensive courses within members’ price range. While there are a few excellent courses, they would like to see more discrete training courses.

2.4 Project Life-cycle Processes and Procedures

The Bancroft owns the copyright of most of the materials in the collections and employs fair use criteria when disseminating the resources. There is a bold copyright statement at the point of access:

“Copyright has not been assigned to The Bancroft Library. All requests for permission to publish or quote from manuscripts must be submitted in writing to the Head of Public Services. Permission for publication is given on behalf of The Bancroft Library as the owner of the physical items and is not intended to include or imply permission of the copyright holder, which must also be obtained by the reader.”

While much of the material is old and can be digitized with no rights problems, there are some sensitive areas that Bancroft responds to as carefully as possible. So far this strategy has been successful and Bancroft continues to monitor this area.

The materials are conserved by the curators in the Library and are routed through the curators during the digitization process. During the papyrus project, the fragile papyri were held in plastic sheeting to ensure no damage occurred. Any necessary conservation is carried out on a project, if material comes to light because of the digitization process. This tends to be opportunistic, where material may be digitized because of its fragile nature, or conservation may be an off-shoot of the project. The image lab takes due care and concern when dealing with fragile materials and curators are present when necessary. Bancroft is discussing access to the original materials and may possibly limit access, but at the moment directs users to the digital version.

Occasionally material is rejected before digitizing; it may sound promising in the catalog but in practice is not sufficiently rich in content or presents physical problems.

The original materials are cataloged by the Library. The digital objects are referenced through this catalog by MARC 856. The digital objects are delivered through the catalog and Bancroft is looking at search tools for the TEI and EAD to be attached to the catalog for more precise searching on the digital metadata. The TEI header holds data about the original and the surrogate and refers to the digital object. At the moment the catalog only allows viewing of the digital object. The finding aids on the OAC central server web page allow searching of the metadata attached to each object. There is a local database for all projects that uses the same format – Access. This database references the catalog. The SGML finding aids are produced from this database for access in the OAC central server.

Where possible, the Bancroft uses controlled vocabulary and thesaurus, e.g. name authority files. The metadata records a little information about the original object, the digital object, the process, technical details, staffing details and administrative information. The metadata is recorded by the digitizers in the image lab. The database outputs the finding aids for web access.

2.5 Access to Digitized Materials

The Bancroft Library mounts some of its projects online through the Californian Digital Archive – the OAC and CDL. These materials are accessed through the SGML finding aids which are produced from the database. The minimum specifications are available and Bancroft adheres to these, thereby enabling its materials to be accessible online.

They are currently researching MOA2, an XML DTD for structural and administrative data. This helps navigate through complex objects – looking at the master and derivative images. The rights information can also be attached by links to the catalog and the finding aids. Output would be as MOA2 objects with the finding aid passing onto MOA2 DTD. This structural metadata allows generic handling of complex hierarchies. The DTD would be in common exchange formats and the Library at UC Berkeley is producing a variety of tools to support the capture of administrative and structural metadata during the creation of digitized archival materials.

“The cornerstone of the MoA II effort is an XML DTD that defines the digital object’s elements and encoding. The project has also developed a relational database that allows a Library to capture the metadata, a program that reads the database and automatically creates the XML encoded digital objects, a repository manager that provides distributed network access to the objects (via RMI) and a viewer that displays MoA II objects from the repository.”[2]

MOA2 is based on administrative data for images; the aim is to record more of the derivatives and the capture process. There is a current concern over the databases as they stand at the moment and problems with migration. MOA2 will therefore address access, preservation and exchange issues.

2.6 Evaluation, Funding and Long-term Sustainability

Bancroft does not carry out any formal evaluation, but have consulted curators and faculty staff at the onset of projects about user needs. There is a perceived lack of standard tools for this process and access to typical instruments would make the process more viable.

Grants from funding bodies often require evaluation and dissemination strategies which motivate these procedures. However, it is also felt that the evaluation process may not always be fruitful and *ad hoc* discussions with colleagues can be as effective in the long term.

As part of the OAC, Bancroft recognizes that updating and adding to the archives is a necessary and unavoidable process. On older projects, the metadata will have to be updated and new materials added. The OAC is looking at changing the user interface, examining off-shelf solutions and may produce a new interface for access. The intellectual effort in MOA2 will help this process. It is through MOA2 that Bancroft is redirecting its long term preservation and archiving strategy so that it no longer relies on the databases. The preservation strategy is based on migration of data; emulation and hardware/software retention are not valid elements. The aim of this strategy is to ensure that the digital objects remain usable for as long as possible as their loss matters greatly.

2.7 Comments and Conclusions

The Bancroft is a university Library that participates in many digitization projects across the campus of UC and across the state. It has an in-house image lab with excellent staff who manage the digitization of a variety of materials. The Bancroft has its own guidelines for internal projects, but is also a major part of larger programs, such as the OAC, and outputs finding aids in EAD for use in the OAC online server.

The Bancroft, in common with many projects and programs in Berkeley, is aware of the need for good, easy to use, cross community tools. It is with this in mind that they are researching the XML DTD for MOA2. The re-addressing of these larger issues is a timely and strategically thoughtful method to ensure the longevity and usability of the digital objects that Bancroft produces.

They are motivated by research from faculty to create their own digitization projects, but their material is also available through cross disciplinary, cross community projects such as JARDA. In being an individual project as well as an integral part of the larger university and state program, Bancroft has ensured that it can serve both the immediate community and the growing numbers of users of digital material, while also ensuring that its collections are accessible and will remain so in the long term.

---

[2] http://sunsite.berkeley.edu/moa2/

valid xhtml 1.1
abp~04/02

3 Brown University, Women Writers Project

On September 19 2000, HATII interviewed Julia Flanders, Director of the Women Writers Project at Brown University. The project has as its long-term goal, the symbiosis of early women's writing and electronic text encoding, thereby making texts written by pre-Victorian women accessible to teachers, students and general readers alike. By the end of the 1980s, researchers at Brown recognized two emerging communities: that of the growing sphere of early women's studies and the developing field of electronic text encoding. By reconciling the two through digitization, the Women Writers Project hopes to improve access to the material, while also ensuring its long-term preservation.

3.1 Organizational Digitization Program and Policy

No processes or activities led to a formal digitization strategy being developed, nor was a collection survey undertaken. However, the original project bid did identify the corpus to be digitized (women's writing in English from 1350 to 1850). The corpus was estimated at roughly 10,000 texts. The digitized collection was initially estimated at about 1000 texts, but this has been revised and the target collection size is now open-ended. At this time (1986) the project was centered around a group of scholars from a range of institutions rather than locally within the institution.

Texts were originally selected for inclusion in an electronic anthology, since the first priority was to make texts available for teaching use, to remedy the lack of early women's writing in print. Subsequent collection development has emphasized broadening the collection in particular areas: the Renaissance period, non-literary texts, religious debate, etc. The original priorities were established by the project scholars; expansion of the collection is now guided by project staff and scholars. The priorities were not formalized in any strategic policy statement but the project guidelines for selecting texts are published on the web.

The objectives that this policy sought to achieve were: firstly, that there should be adequate chronological coverage and generic diversity and, secondly, to respond to scholarly needs. The project believes that it has achieved these objectives but that there were also obstacles — it is the many banal factors that affect digitization. For example, some genres take a long time to digitize (drama), which ends up skewing the digitized body, or the health of working relationships with libraries affects the speed of delivery of texts.

Obstacles to planning the development of digital deliverables included a lack of relevant standards in the beginning and a lack of depth to the organization, for example in the ability to train people consistently. In terms of the process of building digital deliverables, obstacles included the length of time it took scholars to understand TEI and how it represented text. There were also other problems of being an early adopter (P1 - P3) and legacy data pre-1993. In the absence of other projects to model on, it took some time to evolve an efficient structure; initially there were very few project staff and hence problems communicating between the scholarly board and the encoders.

The primary selection criteria for digitization were the materials' historical and cultural value, followed by their research significance and teaching and learning potential. Other selection criteria, that could not be ranked, included improved functionality, enhanced access, the potential to reach disadvantaged groups and social inclusion. It was not felt that any of these selection criteria could be given a prioritization ranking. It was also felt that these selection criteria had changed over time but only in the specifics as the collection developed, for example an increased interest in non-literary materials including science and medicine. Related to this there were also recent changes in prioritization criteria, for example prioritizing texts by women of color and a de-prioritization of manuscripts, long novels, original cookbooks and material already in print, with the exception of considering novels as a test for OCR, and manuscripts as a possible future experiment.

The form of co-operation has been largely institutional within Brown University, including the University Library and with other projects. Recommendations from this experience include the pooling of resources, avoiding duplication of labor, thinking ahead about publication and ensuring co-operating bodies have compatible exit

strategies.

The Women Writers Project started in 1986, with funding from the NEH starting in 1988 and continuing until June 2000, and is ongoing with no anticipated end date.

The digital deliverables were created primarily for wider access, with research, teaching and learning resource and experiment also as significant purposes. A further purpose was public access, whilst revenue generation became a goal of the project. The project had originally produced a statement of intent but the details of this were not known.

The types of source materials digitized include printed books, printed documents, unbound printed documents and handwritten documents. In all cases the nature and format of the materials was photocopies or prints from microfilm. The digitized deliverables from these sources were intended to represent a sample of the corpus, rather than the entire body of material. The deliverables were intended to be re-purposed but in an open-ended way, for example customized anthologies or Braille versions. However, no specific plans exist and in practice this has meant only two or three versions of the material — printouts, online and published books. There have been some difficulties in managing these multiple output formats, largely due to a lack of appropriate conversion tools. As a result a disproportionate amount of time has been taken up with writing and maintaining tools for printing and data conversion.

The Women Writers Project used a number of standards, guidelines or tools for representing content. These included TEI, SGML and XML. For describing content the project used TEI headers and Brown Library has created MARC records to describe the project's materials, which are distributed publicly. For controlling data values the project developed its own name authority files and did not use any other method because of a lack of resources. The project also looked to guidelines for digitizing particular document types and used TEI as soon as it became available, although extensive modifications were made to it. COCOA and MECS were rejected as guidelines because they did not have an adequate level of international support and standardization. For representing structure the project used TEI, SGML and XML.

In relation to standards in general the project's view is that there is no ideal in encoding standards and that a project cannot be, for example, "pure TEI" but must customize to some extent. At this point in time all research is trying to establish an ideal.

The primary target audience is

- four-year college;
- Graduate school.

The secondary audience is:

- distance education.

Distant priority audience (ranked 10) includes:

- K-12
- community college;
- lifelong learning;
- public library users.

Those outside the intended target audience could use the digital deliverables and the most likely group was those beyond the four-year college level. The profile of the actual users has been the one anticipated.

It was not felt necessary to evaluate the needs of the core group of the target audience, as they had provided suggestions for texts. However, a survey of attitudes to the use of electronic texts was undertaken for the Renaissance women section of the project (response of 69 from 300).

The project has acknowledged the needs of those with disabilities through the use of SGML, which means that the website can be used by those with text based browsers. The only limitation to the use of the digital deliverables is minimum system requirements (which are not high) and these are clearly stated.

3.2 Project Management and Planning

Advice on project management was available in-house. Last year the project was absorbed into Scholarly

Technology Group (STG) but only for administrative purposes. Formal project management procedures include an advisory group to the executive committee, which in turn advises the project director, who reports to the director of STG. One less effective management procedure was an organizational structure that was insufficiently hierarchical. This had the result of much undifferentiated work revolving around one person, (although this could also be a strength).

No feasibility, pilot, time and motion or benchmarking studies were carried out. Planning and scheduling tools include the use of a database to track workflow and paper based charts. The use of student workers was based partly on pedagogical motives and partly on reasons of quality, as they had no preconceptions, can be trained effectively, and are intelligent and well motivated. Both job descriptions and performance indicators are used for staff, the performance indicators coming from Human Resources.

Digitization is carried out in-house as the project does not believe there is anyone else who could do it with the amount of content encoding involved, and it is therefore easier to transcribe and encode at the same time. For this purpose the project bought in equipment (Mac computers). The process of typing in texts was chosen because the source texts are heterogeneous and OCR cannot be used efficiently.

3.3 Human Resources and Training

The number of people working on the project and their capacities are based on 1998/9 figures when staffing was at its capacity. There was one FT project director, at 100% on project, one FT textbase editor who was also a metadata specialist at 100%, one FT trainer at 75%, one FT electronic publications editor at 100%, one FT licensing manager at 51% and 20 (10 FT equivalent) encoders (digitizers). The project also shares one FT administrative manager with STG at 50%. All the staff had humanities backgrounds with the exception of some of the encoders, who were all Brown graduate students and learned humanities computing on the job. There were no content specialists on the project staff. None of the staff was redeployed from other areas and advice on the technical aspects of digitization was available in-house.

Training needs were assessed on an informal/incremental basis. Latterly a training package was put together for the trainer on the application of TEI technical standards. Other training areas identified included the technical operation of digitizing equipment (Macs), metadata creation and documentation system. This training was undertaken by the project's trainer and was organized in-house using internal consultants as well as independent study and learning on the job. The training has met the needs of the project.

3.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables it has created and this is declared on the license to use the materials but not on the website. It does not own the copyright of the original materials, which are all public domain. No primary material in copyright was digitized, but secondary material was digitized with the owners' agreement and without formality. Users of the digital deliverables are allowed to make printouts on paper, and under a site license to download to a PC, LAN or WAN. Users are able to download ASCII text files, TEI DTD marked-up text, XML marked-up text and PDF formats. No electronic management system is used to control copying.

Preservation and conservation policies and procedures are not applicable as all material is digitized from photocopies or prints from microfilm.

The cataloging or reference system used before digitization is unknown (materials come from a variety of libraries) but the information that is used in the digitization process (the TEI Header) includes the title, author, size, format, library and anomalies. The project does not have access to all the relevant cataloging information before digitization and early on the project was ignorant of much of this information at the encoding level. As a result the project has had to locate core reference material for the digital deliverables. An additional problem has been the poor documantation of microfilms.

No materials were altered from the originals because all digitization was done from intermediaries (photocopies, prints from microfilm and in the past single microfiche) although it did not just exist in reproduced form. Some material was rejected because of illegibility.

The project does not catalog the original materials and uses an in-house Filemaker database to catalog the digital surrogates. Data values are organized by controlled vocabulary (data value in catalog — author,

translator, text format, anomalies) and a genre, name relationship database.

The metadata records details about the original object, digital object, digitization process, technical details, staffing details and administrative details. The metadata record is created by the encoder (digitizer) from information provided by the textbase editor. The catalog for the digital deliverables is held in electronic form on an intranet and is also available on the internet (website). The original and digital records shared some information and were independent in other respects. The catalog and record are linked through a unique ID for each object, which becomes the ID number for the transcript and TEI header.

3.5 Format, Resolution and Compression of Digitized Materials

The formats for retroconverted digital text are:

- SGML;
- XML.

Some of the text contained non-Latin scripts. Keying-in was used in preference to OCR. The lessons from this experience were that the students became fully involved in the texts and this was valuable for intensive encoding, but needed a responsive coding scheme.

Quality control procedures for the deliverables are proof-reading, in-house SGML validation tools for checking encoding consistency and values, and a review stage. Quality assurance issues are also reported by users. These quality assurance methods evolved in response to staff input, periodic review and refined methods of proof-reading. Metadata are also proof-read, but not online. One effect on workflow, was that to have an encoding review stage after the first proof meant that the proof-readers were having to deal with many errors even if the encoding errors were not repeated.

Users have open access to the catalog and access to the digitized materials is through an annual license. Users can search on word, phrase and proximity, while Boolean searching can be done on the metadata. Searching is context sensitive at the collection, subset and text levers. This searching is handled by the TOC and main text frames, and there is also a summary of metadata above the title page. The project had to design many Dynaweb, IP and password authentication systems as well as a customized interface for each text. Users can save file sets and the project hopes soon to be able to allow annotation shortly. Usage is monitored by automatic data capture and the project will use formal data collection in the future.

Users have to pay for the digital deliverables if they are outside Brown University or the general public. Users are charged annually, pro-rata to the size of the institution, and this charge is collected by invoice. The revenues were estimated in advance and in the first year receipts totalled $139,000 (including advance payment for multi-year subscriptions); this year receipts were $24,000 plus $58,000 for next year. It is not known if the revenues generated met those estimated.

Potential users were informed by website, articles and advertisements in print media, conferences, meetings, email shots and library links and portals. The most effective methods have been conferences and conventional mails, with an email follow up.

3.6 Evaluation, Funding and Long-term Sustainability

No front-end or summative evaluation has taken place. Formative evaluation took place with a beta test group through online questionnaires, focus group discussions and observations of user interactions with the system. As a result the interface was adjusted, especially the navigation labels and what information was displayed at what stage, e.g. the TOC and the behavior of frames.

The project has cost $2,801,113 so far and the major sources for this have been Brown University, NEH, Mellon, private donors and industry. In terms of how much it should have cost, the project estimates that 10% of the funding was “wasted”, but of this 10%, 25% was “wasted” in a good way, in learning about the processes. The project felt that grant proposals were under-funded by 25-50%; if it had this money there would be better delivery and an expanded collection through more investment in programming, training, documentation and bibliographic research. It was felt that the use of standards perhaps saved money in the long-term. The NEH were sent reports every six months for their monitoring of the project, and the Mellon Foundation received several reports, but generally funding organizations have not been demanding. The documentation asked of the

project by funding organizations consisted of interim and final reports.

The project elements being updated and their frequency include: new digitized material (quarterly); new metadata (added annually); changing metadata (tailing off); and the user interface (updated approximately once a year). The project's preservation strategy for long-term access to the digitized objects is the use of SGML. There are no quality control procedures for life-cycle management.

The project intends to keep the materials available indefinitely. The longer-term sustainability is dependent upon self-generating funds and the resource will generate sufficient income for this, although not all of these funds are secured. The project has an exit strategy based on STG administering the data to subscribers. The loss of the digital deliverable would be a matter of concern.

3.7 Additional Comments

The instrument was particularly good for libraries but some areas needed to be addressed with more nuances for some projects.

valid xhtml 1.1
abp~04/02

5 The Art Institute of Chicago, Imaging Department

Alan Newman, Executive Director of the Department of Imaging at the Art Institute of Chicago, was interviewed by HATII on May 11 2001. The Imaging Department is responsible for creating an archive of more than 150 000 images, for the purposes of reproduction, reference and display. The Archive is a coalition of ten curatorial departments from within the Art Institute, and aims to effectively reflect the activities of these departments, by increasing the accessibility of their resources.

5.1 Organizational Digitization Program and Policy

The Art Institute of Chicago's Department of Imaging has developed its digitization strategy by fulfilling basic institutional needs, such as project demands or assessment of the educational value of the material. The technical strategy has then developed through collegial exchange, technology research and membership of organizations like CIMI and AMICO. A decade ago, the department was exclusively involved with the creation and dissemination of high-quality photographs of the museum's permanent collection. Today the department provides an array of electronic imaging, audio-visual, multimedia and data-based services.

Digitization activities first began with *ad hoc*, small-scale projects with focused and perceivable goals, working to short timetables. One of the first projects, in 1992, was to digitize glass negatives for research and preservation with funding from the Samuel Kress Foundation. This provided for electronic cataloging of materials and the birth of in-house digitization, image management, image manipulation and digital video production. In 1993, the Imaging Department produced a five-minute orientation video for the Institute's Centennial Exhibition and the image database provided content for two picture books associated with the exhibition.

For a number of years the computerization of the Institute's permanent collection database, *Chicago Information, Tracking and Inquiry* (CITI) has been underway. Since 1998 the Department, working with the Registrar and Information Systems has been transforming CITI into an Enterprise image database. This will be rolled out to Institute staff in late summer 2001 and the general public in early 2002.

At present the Department is six months into a one-year project to reduce internal reliance on color transparencies and establish a totally digital chain. For now film will still be provided to outside agencies who require it but the image database for internal use will allow people to send print files to either local or enterprise printers on their own. Cost will be reduced from $6-8 for centralized b/w RC photographs which were delivered in one week to 25-50c for 24/48 hr delivery. The Imaging Department can then become a capture, quality control, storage and management facility.

The Institute did not conduct an extensive collection survey as part of planning its digitization strategy. Priorities for digitization are established by a high level adjunct committee to CITI composed of the heads of the Library, Education and Publications. Priorities have not been formalized in a strategic policy statement, but the objectives of digitization are, firstly, educational value, in particular K-12, then scholarly value, followed by good cataloging data being available (for example recently researched or published). These objectives provide the main selection criteria for digitization. Whether the material is also needed for ancillary publications or video is a by-product of these policy decisions.

A recommendation to other institutions who are attempting to formalize their selection criteria would be to have the widest possible audiences represented on the decision making body. For example, the AIC has determined that the widest audience interest would be for images of its paintings including all western and Asian art.

The main overall obstacle to planning the development of digital deliverables has been rapid technological change. Obstacles to the process of building digital deliverables have been staff training, fundraising and recruiting.

The criteria used to guide the selection of material include:

- teaching and learning potential;
- conservation;
- cultural value;
- outside requests from either scholars, publishers or commercial licensees;
- research significance;
- provision of user services;
- enhanced access;
- preservation;
- potential to reach disadvantaged groups;
- infrastructure cost reduction.

These criteria have not changed over time but the pendulum has swung from scholarly research to education as the number one priority.

The Imaging Department has co-operated with museums, academic institutions and foundations and charities at local, regional, national and international levels in developing their digitization program. Most of these collaborations have been informal and AMICO has perhaps been the most important, although Mr. Newman has lectured widely at professional conferences, such as Museums and the Web, ICHIM, Visual Resources Association and the Museum Computer Network.

The primary purposes in creating the digital deliverables reflect the Department's selection criteria. Digital deliverables are primarily an educational and scholarly resource, with factors such as preservation, access and outside requests as concomitant factors.

The type of source material digitized includes:

- printed illustrations;
- original artworks;
- color transparencies;
- photographic negatives;
- photographic prints.

The predominant mediums of the materials digitized are 4x5" color transparencies and original art (paintings have been captured on color film since 1985). The digitized deliverables represent a sample of material rather than an entire body. For the Imaging Department re-purposing the digital deliverables has been one of its original aims as a service bureau. Exhibitions and publications are what drive the Art Institute and most images are by-products of these. Despite the serendipitous nature of the digital images this approach has been extremely successful with the Department, in collaboration with Museum Education, winning three national first prizes for its educational multimedia.

The following standards, guidelines or tools are used for representing content:

- CIMI (AMICO);
- JPEG;
- others (TIF, QT and QT-VR).

The following standards, guidelines or tools are used for describing content:

- Dublin Core (AMICO sub-set);
- Categories for the Description of Works of Art (Art Institute of Chicago variation).

The following standards, guidelines or tools are used for controlling data values:

- AAT;
- name authority files (Union List of Artists Names, Getty Thesaurus for Geographic Names, plus standard reference works);
- others (AMICO Object Type List).

The Department looked at other existing guidelines for digitizing particular document types when planning its digitization strategy, most significantly, Making of America and AMICO. Mr. Newman played an important role in the formation of AMICO's imaging practices. The Department adopted QuickTime for motion pictures and QT-VR for gallery and other architectural views. They have also used QT on the web in a new way, to produce a freely available, non-licensable object viewer with panning and zooming attributes. See http://www.artic.edu/cleo for an example.

The following standards, guidelines or tools are used for representing structure:

- HTML;
- JPEG;
- Others (QT, sometimes GIF).

The HTML used for representing structure is not yet Dublin Core or XML compatible, but it could be developed.

The digital deliverables created by the Imaging Department are currently only available in small samples to the public audience. CITI/Image will move many thousands of images as a client-server product in the third quarter of 2001 and publicly on the web in 2002. In July 2001 the Imaging Department will launch an experimental project for licensing images on its website which will include keyword subject query to almost 1,000 images. See http://www.artic.edu/aic/rights.

5.2 Project Management and Planning

Internal advice was available on managing the Department's digitization program. As the Department is now going totally digital some inroads are being made in cross training and walls between departments are coming down. Ultimately there is a need to develop interdepartmental co-operation and services. Unanimous agreemetn among curatorial, conservation, registrar, education, publications, graphics, imaging and library departments on the revolutionary potential of the image database, indicates the future development of this trend. Outside help was sought on color management, while hardware vendors often lent equipment for the Department to trial.

Feasibility studies are being carried out for the purposes of technical feasibility, workflow analysis and technology forecasting. These studies have been ongoing and organic and will be completed by the end of 2001. The Department has not carried out any time and motion or benchmarking studies.

Digitization has been carried out in-house so as to exercise more control over image quality. Equipment has been specifically purchased for the purposes of digitization. Many digitization processes are used, but the determining criterion is that which provides the best quality return for the monetary investment. The digitization equipment the Department uses includes a Scanview F10 for film scanning, a Nikon D1 series digital cameras and a Phase One PowerFX professional camera. Guidelines for data capture procedures have been established. Benchmarks used for image digitization are IT 8 charts and color management hardware and software to produce a closed loop between monitors, scanners, cameras and printers. This hardware includes X-Rite colorimeter, Greitag Spectralino and ColorPar's Optical as well as Apple's Colorsync. The department is also experimenting with MTF targets.

5.3 Human Resources and Training

The department employs twenty people, ten of who are working full time on aspects related to digitization. Four people are employed on image capture and three on post capture processing. Others also work on outputting, technical development, maintenance and metadata. Staff members have a background in photography. As the Department will have subject (content) based access to its image database, a graduate library student will also be employed to check metadata and descriptions within the database. With the exception of color management, expertise on the technical aspects of digitization was available in-house.

Training needs are assessed informally, however the department strongly believes in developing redundant skill sets and cross training. Areas of training that have been identified include:

- technical operation of digitizing equipment;
- post digitization processes e.g. editing.

The team members that have received training include:

- specialist technical staff;
- equipment operators;

- photographers;
- lab technicians.

Training has been organized using the following:

- in-house, using project staff;
- learning on the job;
- seminars offered by the trade.

The training has met the needs of the project.

5.4 Project Life-cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and owns most of the copyright in the original materials. Other copyright holders are the ARS, Vaga, artists and estates and their representatives. The copyright status of the final digital deliverables will be declared. Copyright is currently placed on all material coming out of the Institute and there are future plans to use an embedded visible electronic watermark. The Department will ensure that this is effective, as it does not want to become a policeman. Educational (fair use) users are allowed to make low resolution watermarked printouts of the digital deliverables on paper.

For text material users can download:

- ASCII text files;
- PDF.

For digital image material users can download:

- thumbnails;
- lower quality images;
- associated documents.

For digital image material, users can only view highest quality images. Wavelet compression, JPEG 2000 and pan/zoom server technology are being evaluated.

For digital moving images users can download and view samples.

The Art Institute of Chicago has a conservation procedure for original works of art. The Institute's conservation department routinely discusses with the Imaging Department, strategies for protecting material to be digitized. Equipment and processes for vulnerable materials include special cradles, lighting, exposure limits and anti-vibration measures. Materials are prepared either by conservation or Imaging staff during digitization. Particularly when there is new equipment conservation staff will set lighting, UV, and temperature levels, as it did when the Phase One camera was installed. Access may be restricted to original material after digitization.

The Art Institute of Chicago's in-house collection management system, CITI, is used to catalog the original materials. The CITI system is a client/server database using the cross platform database, 4D. CITI currently contains 130,000 records.

So far no material has been rejected for digitization. For the Institute's own works the Department digitizes from originals for works on copystand or from intermediaries for paintings and 3D objects. The department expects to employ all direct digital capture by the end of 2001; when capture of external material is required the Department has to use whatever it is sent.

In early 2002 the Department will supervise a satellite digital studio in the Japanese print storage area for the purposes of systematic digitization of the Art Institute's distinguished collection of Japanese woodblock prints. It is expected that a high-end electronic flash based digital camera will be selected due to conservation issues.

Intermediaries used for image digitization include slides, 35mm or 4x5" transparencies.

The intermediary used for moving image digitization is Betacam SP.

Modified MARC is used for cataloging the digital deliverables.

Tools used for controlling data values are:

- controlled vocabulary;
- thesaurus;
- classification scheme (in-house based on AAT);
- others (Union List, etc.).

Metadata details are recorded about:

- the digital object;
- the digitization process;
- technical details;
- staffing details;
- administrative information.

The Imaging Department has adopted a conservative metadata standard (AMICO minimum). Metadata records such as iconography visible in the view or detail could be augmented by the curator. Most metadata information is automatically generated by applications. Manually added metadata mainly relates to the image view. The metadata records are held in the Imaging Department's database written in Helix RADE, and will be available to the new CITI/Image system. The relationship of the records in the two catalogs is the same. Placeholders referencing hard camera copy hold gaps in the Enterprise image database. The metadata will be used to control whether the image is publishable or is simply a quickly made document (snapshot). The record and object are linked through the image number.

5.5 Format, Resolution and Compression of Digitized Materials

The TIFF file format is used for the capture, preservation, printing and delivery of image material. JPEG is used as an additional delivery format for web and screen display. The capture, preservation and delivery resolution depends on the size of the original.

For 4x5 or smaller originals the archive image is 30-50MB, for 4x5 to 8x10 originals the archive image is 60MB, for 8x10 to 11x14 originals the archive image is 70MB and for 11x14 and above originals the archive image is 80MB and above in size. Capture bit depth is 16 bits per channel (48 bit) and preservation and delivery 24 bit. It is the 24 bit TIFF file that is saved and the Department does not capture as high as some other institution's projects because the lower levels fulfil 99% of use. JPEG image compression is used for delivery to improve access and enhance usability. The project retains the original scans in uncompressed form. The project carries out post processing on images using PhotoShop 6 which offers a lot of tools for the high bit images. ColorQuartet is also used (scanner software). Processes carried out include gamma correction and sometimes color correction when comparing screen to transparency or object. The dynamic range of the equipment is checked. For others starting work in digital imaging, the Imaging Department would recommend starting small and prototyping.

Digital moving images are created *ad hoc* on a project by project basis. In all cases QuickTime is used for capture, preservation and delivery with resolution set for the best compromise between speed and quality in each case. The Media100 system is used for video editing and is augmented by Adobe After Effects and Terran Movie Cleaner.

The quality control procedures in place for the digital deliverables are carefully calibrated input devices and color evaluation post capture by senior Imaging staff. Metadata quality control procedures are being finalized for content related entries. Field and name controls and some automatic checks will also be used.

Access to the digital deliverables and catalog is restricted to in-house users. Searching and browsing facilities are keyword and Boolean with an emphasis on authority fields. It is too soon to anticipate the level of use with the new system but the Department expects exponential growth.

5.6 Evaluation, Funding and Long-term Sustainability

The Department of Imaging's digitization program has cost $400,000 over 10 years in actual grant support and perhaps an equal amount in staff in-kind. The main sources of funding have been grants and private foundations. If the Department had been able to determine its own funding it would not have done anything substantially different. The Department's view is that the use of standards has saved money in the short and long term.

New material and its associated metadata will be added at the rate of 10,000 items a year and metadata is changed frequently as the nature of collections change. The Department of Imaging's archive strategy is based on a rotation of off-site location, redundant media and optical media (proposed DVD-R when ROI is realized). Quality control measures are in place for life-cycle management.

The Department intends to keep the digital deliverables available indefinitely. The program is dependent on self-generating funds and the creative efficiencies provided by the technology. The Department has not needed an exit strategy for its digitization program because it is in use in practice for general operations within the institution and is integrated into it.

5.7 Conclusion

The Art Institute of Chicago's Imaging Department is notable as one of the few, if only, projects surveyed that has developed a totally digital chain from capture to delivery. In particular the integration of this service with the institution wide CITI/Image system is worth highlighting. This has reduced the cost and time of delivering image material and enabled the Imaging Department to re-orientate its activities away from serving internal users directly and towards image creation, management, tracking and storage. The implications this process has for the whole institution seem no less significant. That the Imaging Department is largely an internal "on-demand" service has meant that it has not been encumbered with some of the strategic selection, prioritization and collection development issues that face other digitization projects. Furthermore, it remains to be seen what impact the public rollout of this system has on workload and orientation. It is possible to envisage that the demand created by the system may pose some strategic digital collection development issues.

valid xhtml 1.1
abp~04/02

6 University of Chicago, The Oriental Institute

On January 12 2001, HATII interviewed Charles E. Jones, Research Archivist-Bibliographer at the University of Chicago's Oriental Institute and John C. Sanders, Head of the Computer Laboratory. The cultural and historical value of the materials held by the Institute was a significant factor in the decision to digitize the collections. To this end, the process would ideally enhance public access as well as most importantly, increasing the potential of the material for teaching, learning and research purposes. The work of the Institute primarily revolves around research, and a considerable shift of focus was required of those employed at the Institute, in order for the digitization program to successfully become a reality.

6.1 Organizational Digitization Program and Policy

The Oriental Institute has a long record of involvement with electronic and digital media. Experimental – and expensive – work was undertaken with entering texts into a mainframe computer in the 1960s. PCs have been used since their introduction in the 1970s on the initiative of individual faculty members. An organized institutional effort began in the early 1990s with the connection of the Institute's building to the local area network. It was then that the Institute decided to purchase a PC for all its members of staff. Subsequently, the convenience of email and printing convinced people to use their computers.

However, the Oriental Institute's digitization strategy was not directly related to infrastructure developments, even if these were a pre-requisite. This began in 1992 when Institute staff spoke with the University of Chicago's Computing Service and they realized that they could serve redundant digital material. The Institute has run a web service since 1994 and it is operated under the assumption that the service will be available indefinitely. However, the Institute does not have a formalized, strategic digitization policy. Rather, the Institute's Research Archivist-Bibliographer, the Head of the Computer Laboratory and the Director of the Institute make such decisions. The interviewees were acutely aware of the need to institutionalize their digitization operations and avoid the yo-yo effect of periodic centralization and decentralization. However, any initiative for this is likely to come from the museum, rather than the research side of the Institute. Another possible locus for a digitization strategy may be the publications department. The Institute has not conducted a collection survey as part of its digitization program and priorities have not been formalized. However, an original principle that was used to support the idea of publishing material online was that it should have been published in print before it appeared on the website. Today many components of the Institute's website have no print analog. The museum's closure for refurbishment in 1996 was an opportunity to photograph the objects but this opportunity was not taken for fear it would slow down the renovation.

The main obstacles to planning and building the development of digital deliverables have been money, manpower and the lack of an institutional overview, although there has never been a philosophical objection to digitization.

The primary selection criteria used to prioritize material for digitization are teaching and learning potential and research significance. This is closely followed by the material's cultural and historical value. Enhanced access was a primary criterion for selecting material when the Institute's museum closed for refurbishment. There is also theoretical institutional support to digitize for preservation and the preservation department is fully supportive of this aim. The driving force behind the Institute's digitization activities are funded projects.

The Oriental Institute has co-operated with museums, academic institutions and professional societies at local, regional and national levels in the development of its digitization projects. For example, the museum's education department is part of a project to deliver curricular materials over the web to Chicago schools. Based on this experience one observation is that collaboration often encourages institutional hesitation and it is vital to get beyond talking to action. It is also pointed out that projects, even when using the same people and working to the same deadline, can have different rates of progress.

The primary purposes in creating the digital deliverables are as a teaching and learning resource, research, and

to increase accessibility. The Institute has produced explicit statements of intent for some aspects of project work.

The type of source material digitized includes:

- printed books, serials and similar bound material (published);
- printed documents (gray literature);
- unbound printed documents;
- typescript documents;
- printed illustrations;
- artifacts;
- original artworks;
- intermediaries;
- hand copies;
- photographic negatives;
- photographic prints.

The predominant medium of the materials to be digitized is paper, but some velum has also been digitized. Some out-size material has been digitized by a vendor. The digitized deliverables have usually represented a sample of material rather than an entire body, but the Institute may digitize a whole season's field drawings or photographs, for example. It is the Institute's intention that the high-resolution material be re-purposed. For example, digitized vector lines from a map were built to provide a digital model on video on the web, but were also used to create architectural plans.

The following standards, guidelines or tools are used for representing content:

- TEI;
- Unicode;
- JPEG;
- MPEG;
- XML;
- Others (TIFF, EPS, and GIF).

Modified MARC is used for describing content.

An in-house library system is used for controlling data values. Other standards, guidelines or tools for controlling data values were not appropriate to the projects.

The program looked at other existing guidelines for digitizing particular document types when planning its digitization strategy. These included XML, and other DTDs being developed to either use or modify. Guidelines were rejected as being unsuitable for archaeologists and philologists.

The following standards, guidelines or tools are used for representing structure:

- HTML;
- Unicode;
- JPEG;
- XML.

The Oriental Institute is still in the process of determining how best to navigate between the ideal and the realistic in the application of standards. The Institute is trying to develop a DTD and deal with the overriding difficulty of how to represent scripts for which there are no standards.

The target audience for the digital deliverables are four-year colleges and graduate schools, but very much aimed at the teachers. Similarly, there is a current drive towards the K-12 audience, but again aimed at teachers through an ongoing project in the education office. The general public museum users are also a target audience, through the Institute's general public display collection material rather than the more tailored or specific material geared toward other audiences.

The Oriental Institute's education department has undertaken an evaluation of its target audience using a 40-member focus group of high school teachers. This was a full written evaluation and was a requirement of the funder. Aspects of the material were revised as a result.

The Institute has not explicitly taken account of the W3C's web accessibility guidelines. However, they are

represented on an accessibility group that discusses this and works out of the University's networking services division.

Limitations are placed on some digitized material; these restrictions take the form of password restricted areas of the website, users needing to be enrolled in a particular class, and some material is for in-house use only. In all cases these limitations are in place because of copyright or IPR restrictions.

The profile of users has been surprisingly broad in terms of country and the types of files used.

6.2 Project Management and Planning

Internal advice was available on managing the Institute's digitization program. The management of the digitization program has fitted into the existing departments and units of the library and computer laboratory. The digitization program has not led to any changes in organizational relationships and procedures although there is a degree of push and pull between departments.

Formal project management procedures are embodied in the Institute's committees, one of which is computational and makes recommendations. The advisory board for publications also makes recommendations although it has an *ad hoc* existence. However, these have not been terribly effective policy making bodies. There are no managerial quality assurance procedures in place and staff have worked personally in an entrepreneurial way without formalized structures.

Several of the projects undertaken were conceived as pilot projects in themselves. The education office may have undertaken feasibility studies in their own right but the details are not known at present. No significant changes to the design of projects has resulted from these pilot studies. The Institute has not carried out any time and motion or benchmarking studies.

Work is allocated on an *ad hoc* basis depending on staff availability. Some staff have been hired specifically for digitization work; however, graduate students and especially volunteers have carried out the majority of work.

Digitization has been carried out in-house because of the cost and risk of transporting the original materials and costs of outsourcing. Equipment for digitization has been bought in and the Institute has tried to forecast its technology needs in purchasing equipment. The decision of which digitization process to adopt is made through discussion of the outputs required, equipment available and the nature of the original material. The equipment used for digitization is Apple, Epson, HP and Umax flatbed scanners, Nikon film scanner and Nikon and Olympus digital cameras. A lot of second stage digitization is done from high quality photography.

Data capture guidelines are in place but not for projects in the field, where the conditions make this impractical. However, the field station equipment is cleaned and calibrated before it goes out and includes capture instructions. Grayscale and color charts are used in the photographs as benchmarks so the color can be adjusted.

6.3 Human Resources and Training

The Institute's digitization program employs one full-time director (Head of Computer Lab) who also undertakes technical support and development. One part-time digitizer is employed, plus a number of graduate students and volunteers who work anything from a few hours to a full week. A consultant is also employed who devotes approximately 25% of their time to digitization projects. The graduate students employed have a Near Eastern background whilst the volunteers have some intellectual interest in the subject. In-house expertise is available on the technical aspects of digitization and the University of Chicago has staff with a wide variety of expertise, for example in panoramic movies.

Training needs are assessed informally. Areas of training that have been undertaken include:

- application of technical standards;
- preparation and handling of materials for digitization;
- technical operation of digitizing equipment;
- post digitization processes e.g. editing.

One area where further training is needed is metadata creation.

The team members who have received training include:

- project director;
- specialist technical staff;
- equipment operators.

Training has been organized using the following:

- in-house, using project staff;
- learning on the job.

The training has met the needs of the project.

6.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and owns the copyright in almost all the original materials. The copyright status of the digital deliverable is declared on every web page. In addition the Institute has a general statement of copyright and permissions linked from the banner navigation tool in its website. With the exception of a few isolated abuses this practice has been effective. Material in copyright was digitized with the owners' agreement for which written permission was obtained. Users are allowed to make printouts of the digital deliverables on paper and film, burn to a CD and download to a PC, LAN or WAN providing they do not republish without permission.

For textual material users can download and view HTML files, with one exception of a text database that creates HTML on the fly from queries.

For digital image material, users can download and view:

- thumbnails;
- lower quality images (these are 4½-inch high intermediate images for low-resolution small-scale monitors).

For digital moving images users can only view QTVR clips.

No electronic management system, such as watermarking is in use.

The Institute's museum staff approve both the material and the process prior to digitization. Risk assessment and cleaning are undertaken prior to photography and curatorial staff make the final recommendations regarding risk. Curators monitor and handle materials during digitization to minimize risk; light limits are set and special cradles built. Once material has been digitized no access restrictions are placed on the originals. The museum may send the digital surrogate instead of the user having to visit in person; they may also serve the digital collection first to users who visit.

Some items in the museum collection are uncataloged, on paper or card catalogs or in electronic records; however, all the library material is now in an electronic catalog. If no catalog record exists prior to digitization the handlist is processed as the basic electronic catalog entry. Sometimes project staff have to locate reference or source material and this is a continuous and developing process and the electronic catalog is updated.

No material is altered from its original state for digitization. Material has been rejected because scanning may cause damage or for procedural reasons. Intermediaries have been scanned, although the material did not exist only in this form. The collection index is one example where the material was to be displayed on the web and quality was less of an issue.

Intermediaries used for image digitization are:

- slides, 35mm or 4x5" transparencies;
- photographic prints.

Original materials are cataloged in the library electronic catalog, with each item accorded a separate entry using the fiche caption, not that affixed to the print. Some original material in the archive is not recorded in the electronic catalog.

Modified MARC is used for cataloging the digital deliverables.

An in-house classification scheme is used for controlling data values.

Metadata details are recorded about:

- the original object;
- the digital object.

Metadata records are created by the digitizer and then checked by an archivist/information professional. The metadata records are included in the main library catalog but the Institute does not have a unified catalog to include material in the archives. This library catalog is available on the internet. The records for the digital deliverable and the original digitized material are the same. The question of how the catalog and the objects should be linked is still under development.

6.5 Format, Resolution and Compression of Digitized Materials

The formats for retroconverted text-based digital deliverables are:

- HTML;
- XML;
- PDF;
- txt (ASCII);
- word processed;
- database;
- Postscript;
- EPS.

Some documents contained non-Latin scripts. OmniPage was the OCR software mainly used, but also TextBridge. Accuracy levels varied with print quality and font, but 90% and above was achieved. Pre-OCR treatment included enlarging and increasing the contrast. The aim of OCR was enhanced searching. Recommendations from this experience would be to increase the text of the source as large as it can go and ensure the OCR software has the right dictionary. Keying in was also used and the recommendation from this experience would be to key it in twice and carry out file comparisons.

For image material the TIFF and PICT file formats are used for capture and preservation depending on the end goal. GIF, JPEG, Postscript and PDF file formats are used for delivery. The capture and preservation resolution is 300-600dpi or 1200dpi; 2400dpi is used for slides. Delivery resolution is 72dpi. Capture, preservation and delivery bit depth is 24-bit for color images. JPEG level 3 or 7 and GIF compression formats are used for delivery to improve access speed. The program retains the original uncompressed files. Post processing operations using PhotoShop include cleaning and sharpening so long as they do not detract from the level of faithfulness to the original. The average file sizes vary but are approximately 5-9MB for archive TIFFs, 300-500KB for 8x10 images, 80-100KB for 4½-inch intermediary images and 18-25K for thumbnail images.

For others starting work in digital imaging the Oriental Institute would recommend considering carefully the trade off between scanning for the immediate purpose and scanning archival TIFF files once, in terms of the ramifications on disk space, money and re-scanning at a later date. A further recommendation would be never to touch the archival copy, and for any work whatsoever to be done on a second copy.

The only quality control procedures in place for the digital deliverables are visual checks although the Institute recognizes the need for better editorial policy on the deliverables. Metadata quality control procedures are more than one check on each record.

Users have a variety of access rights depending on their permissions, including open access to the catalog, open access to the catalog plus materials, and access restricted to in-house users. Users are also able to zoom in on QTVR movies. Searching and browsing facilities are keyword, Boolean, field operators and proximity. Metadata searching is handled in the same way. The level of usage is 300,000 hits per week on the website and usage goes up approximately 20% pr annum, with additional sharp seasonal increases at the beginning of the academic year. Usage is monitored by automatic data capture.

6.6 Evaluation, Funding and Long-term Sustainability

The Oriental Institute carried out front-end evaluation of users for its web page front end using small focus group discussions.

It is impossible for the Institute to estimate how much its digitization activities have cost, but it is less than imagined and the value of the work in relation to the effort and cost should not be underestimated. The education office has external private foundation funding; otherwise sources have been a mixture of government and non-governmental organizations with a significant amount of funding coming from charities for specific components. If the Institute were able to set its own funding, then ten years ago a very worthwhile activity would have been to develop a unified database for all projects. The Oriental Institute's view is that the use of standards will allow it to move to future formats and it does not perceive the benefits in monetary terms, although projects have anyway tried to avoid using experimental formats. Funding organizations have requested project management plans, cost models, and workflow reports as pro forma items in bids.

New material and metadata are not added to a regular timetable, however, a major update occurs annually. The user interface has been changed once but may change annually in the future. The Oriental Institute does not have a formal strategy for digital preservation to ensure long-term access. At present text and image material are archived in the same way on removable drives and CDs, and the website drives are RAID protected. There is no off site storage or backup. This archiving/preservation strategy is based on migration of data and there are quality controls in place for life-cycle management.

It is intended to keep the digital deliverables available indefinitely. Longer-term sustainability is dependent on self-generating funds although verbal assurances have been received regarding the web publishable material. The project does not have an exit strategy if the Institute could no longer afford to sustain the digital deliverables. In this eventuality, the relationship between the Institute and the University would be critical. Loss of the digital deliverables would be significant.

6.7 Conclusion

The Oriental Institute at the University of Chicago provides an excellent example of the problems facing organizations in developing a strategic view of digitization when a significant proportion of their activities are research based. Where there is significant research digitization it would appear that there is little or no motive for the researchers concerned to adopt the same standards or view their output as a deliverable for a wider audience. On the other hand, the Institute's "public face", in the form of the museum and publications department do have motivation and a need for policies and procedures to guide the selection and prioritization of material. Add to this the fact that the driving force behind digitization is but a few individuals, and it is extremely difficult to gain sufficient momentum across the organization for an "institutional view" on digitization. As the interviewees rightly pointed out, the motivation for this is likely to come from the museum and publications department and it may be that procedures are formalized for these areas of the Institute's activities while research based projects continue to operate independently. The Oriental Institute also highlights the problem that developments such as TEI DTDs do not yet adequately address more esoteric material. Whilst the guidelines such as the TEI have user extensibility built into them, this would not appear to be sufficient in this case. Therefore, if one is making a recommendation in relation to standards of "adopt, adapt and emulate" some projects must recognize that this may not be possible for their material, or that they will suffer redundancy or reformatting.

valid xhtml 1.1
abp~04/02

7 University of Chicago Library

HATII interviewed Kathleen Arthur, Head of Preservation Reformatting and Digital Initiatives at the University of Chicago Library, and Elizabeth M. Long, Co-director of the Digital Library Development Center, on January 9 2001. The digitization activities at the University Library have resulted in the creation of several individual programs and projects. While these projects are designed to satisfy specific requirements, the Library also has the general intention to broaden the availability of resources to audiences other than University Library users, by means of Internet access. Simultaneously, they hope to encourage resource-based leaning and the use of technology in education and research both within and out-with the University community.

7.1 Organizational Digitization Program and Policy

To date the University of Chicago's digitization activities have been project based. Initially the preservation department began with a microfilming grant that included a small grayscale image capture element, although there was no OCR used at this stage. At the same time the library as a whole had an interest in developing these areas. The library did not conduct a collection survey as part of its digitization planning process, but a condition survey was undertaken ten years ago. This survey is not used directly to establish priorities for digitization but it raised questions and helped develop experience and understanding that in turn informed digitization priorities. These priorities have not been formalized in a strategic policy statement but one is currently under development. The extent to which the whole institution and library staff participate in digitization differs between projects. The preservation and systems department are always involved and frequently the subject bibliographer.

One obstacle to planning the development of digital collections identified by the University of Chicago Library has been staffing. As areas of new expertise have been required more staff have needed to be recruited. The greatest obstacle to the process of building digital collections has been the question of what to select for digitization. This is why a policy for implementation and coordination is under development.

Intellectual property rights are the make or break criterion for the selection of materials to digitize as the Library has few resources to deal with this. Providing there are no intellectual property issues the materials teaching and learning potential, historical and cultural value and research significance all have a high priority. Enhanced access and improved functionality are secondary, but concomitant, priorities. Of particular significance is the notion that digitization adds something extra, for example in the case of microfilm. Research into digitization strategies and preservation also receive a high priority but on a large scale or collection level.

For the reformatting department these criteria have not changed over time, although the Library's electronic reserve would save staff for other duties.

In developing its digitization program the University of Chicago Library has co-operated with archives, academic institutions and foundations and charities at local, national and international levels. From this experience, a recommendation for other institutions thinking of starting collaboration would be to fully specify everything in advance – for example, what each partner would contribute – while at the same time remaining flexible.

The Library's digitization program began formally in 1994, although projects such as ARTFL had existed prior to this and there had been other informal activities. These digitization activities are ongoing with no anticipated end date.

The digital deliverables are created for the purpose of a teaching and learning resource, public access, wider access and preservation. The Library has also tried to anticipate demand rather than respond to it. The Library produces explicit statements of intent for digital proposals that cover the rationale, scope, significance, primary audience, long-term sustainability, level of faithfulness and suitability for different target audiences. Future developments include putting these statements on the web.

The type of source material digitized includes:

- printed books, serials and similar bound material (published);
- unbound printed documents;
- typescript documents;
- printed illustrations;
- hand-written documents;
- intermediaries;
- mechanical copies;
- photographic negatives;
- photographic prints;
- sound recordings.

The predominant medium of the materials to be digitized is paper based, but the Library has also digitized 4,500 glass lantern slides. The digital deliverables represent both a sample of collections and entire bodies of material depending on the project. It is the intention that the digital deliverables can be re-purposed. The Library currently maintains a website for K-12.

The following standards, guidelines or tools are used for representing content:

- JPEG;
- SGML;

The following standards, guidelines or tools are used for describing content:

- MARC;
- Dublin Core;
- EAD;

Intermediaries (i.e. a second generation medium derived from the original) used for image digitization are:

- AAT;
- Library of Congress subject headings;
- Library of Congress subject thesaurus;
- name authority files;
- other: A Glossary of Geographic Terms, Encyclopedic Dictionary of Physical Geography.

The program looked at other existing guidelines for digitizing particular document types when planning its digitization strategy. These included the Cornell benchmark, UVA, Making of America and NARA. The Library modified these guidelines as required by adding more fields, for example local Dublin Core fields.

The following standards, guidelines or tools are used for representing structure:

- HTML;
- JPEG;
- SGML.

In relation to standards in general and navigating between the ideal and the realistic, the University of Chicago Library suggests it is critical to look to the future and not to subvert standards for convenience.

The primary intended audience for the digital deliverables are the university and research community (four-year college and graduate school). However, there is also secondary interest in K-12 and lifelong learning. The Library is also interested in public, specialist, government and private sector users, as the material becomes available to them.

An evaluation of the K-12 audience has been undertaken as part of the eCUIP project. This involved participation by public school teachers. The Library's digital projects have not formally acknowledged the W3C's guidelines on web accessibility but have adopted a standard approach and the use of conservative HTML to keep our sites available to as wide an audience as possible. The University of Chicago retains rights on all digital deliverables and permission is required for re-use. The profile of University of Chicago users has been as anticipated, but there have been a higher than expected number of international users and a wider range of uses for the library's environmental photographs collection.

7.2 Project Management and Planning

Internal advice was used for managing the digitization program. The management of the Library's digitization program is integrated into the Library's structure. In terms of preservation, this has meant a change in their type of activity rather than any fundamental re-organization. Developments in procedures have seen the establishment of the Digital Library Development Center to coordinate the Library's digitization work as more staff have been devoted to digitization activities. Some of these activities have now been centralized, such as metadata and cataloging, while others such as web authoring are decentralized. The Library believes institutions may need to evolve new structures but consider a named project manager to be essential. Several new groups have been created to aid digital development in the Library.

Managerial quality assurance procedures include project timelines, reports and accountable structures.

The Library has carried out very small-scale pilot studies as part of grant writing for scheduling and technical feasibility. Because of the early undertaking of these pilot studies no significant changes have been made to project designs. The Library has found that it is other factors, such as technology that cause changes in projects. The Library has also undertaken some benchmarking studies for other groups, again as part of external grant writing. Who does what work on a project is determined by staff expertise and availability and on some occasions a new post is created.

So far all digitization has been carried out in-house, using equipment bought in for the purpose with the exception of a very small amount of specialized material (e.g. reel to reel). Which digitization process to adopt is determined by the needs of the original material. Equipment used in digitization includes Agfa Ultra Horizon and Epson 816 flatbed scanners and a Minolta PS3000 scanner with grayscale output. Guidelines for data capture procedures have been established. They are 300dpi 8-bit grayscale, 600dpi 8-bit grayscale and 300dpi 24-bit color. Benchmarks selected for image digitization are: Kodak grayscale Q13 and color separation guide, Kodak Q60 color input target and an Aimm scanner textchart No2.

7.3 Human Resources and Training

The Library employs two digital development co-directors (2 FT equivalent); full texts are now parsed by a vendor. Technical support and development staffs equate to 2 FT equivalent. Within the digitization unit the Library appointed one person for scanning and data input for the first project. They now have one FT manager of the digitization unit with an additional 1-3 FT equivalent students for capture and data input.

Advice on the technical aspects of digitization was available in-house. There are no formal methods for assessing training needs. Training areas that have been identified include:

- project management;
- metadata creation (in particular Dublin Core).

The team members that have received training include:

- project director;
- specialist technical staff;
- library/cataloging staff.

Training has been organized using the following:

- in-house, using project staff;
- attendance at external courses;
- independent study;
- learning on the job.

The training has met the needs of the project.

7.4 Project Life-Cycle Processes and Procedures

The Library is aware of the copyright position of the digital deliverables and owns the copyright in some of the original materials. Copyrighted material is digitized for electronic reserve but is only available on the web to enrolled students and is not intended for long-term preservation. Where material is in copyright, the owner has been paid a fee. The copyright status of the digital deliverable is declared in a condition of use statement on the

web site. Users are allowed to make printouts of the digital deliverables on paper and download them. For non-e-reserve digital materials, users can order high resolution versions delivered on CD.

For textual material users can view:

- PDF (electronic reserve materials);
- JPEG page images.

For digital image material users can download and view:

- Thumbnails;
- Reference quality images.

For digital audio material users can listen to:

- Full length compressed streaming 2-3 minute pieces.

Minimal electronic management systems are in use.

The University of Chicago Library has a conservation procedure for original materials. Prior to digitization investigations are made into the binding (binding pressure), paper quality, size of material, operability and legibility. Conservation activities are built into the project workflow and include encasing in Mylar, disbinding, rebinding and boxing. A book's structure is risk assessed prior to and during digitization and monitored by curatorial staff during digitization. Book cradles are used to minimize risk and sometimes material is rejected because of incompleteness, blurred pages or fragility. Once material is digitized there are no restrictions placed on access to the originals.

A full range of cataloging systems is in place before digitization, ranging from an inventory to a full electronic catalog. All the available information is used from these systems in the digitization process. Where a project has to locate additional reference material the key criterion is how much metadata would allow the material to be usable.

The Library has digitized from intermediaries, but the material often takes other forms. For example, sometimes only photocopies are used because the originals are too large and need to be reformatted.

Intermediaries used for image digitization are:

- photocopies;
- slides, 35mm or 4x5" transparencies;
- photographic prints.

Original material is entered into the online catalog or a collection level record is created. If there is a record for the original collection material then one is created for the digital collection, but enhanced with, for example, a description for each image in a database.

The following standards or guidelines are used for cataloging the digital deliverables.

- Dublin Core;
- USMARC;

The Library has developed its own in-house database using Microsoft Access or MySql.

Tools used for controlling data values are:

- controlled vocabulary (LC subject);
- thesaurus;
- classification scheme.

Metadata details are recorded about:

- the original object;
- the digital object;
- the digitization process;
- technical details;
- staffing details;

administrative information.

The digitizer and an archivist/information professional or the digitizer who is also an archivist/information professional creates metadata records. Collection level metadata records for the digital deliverables are included in the main catalog and sometimes in a separate catalog and are available on the internet. The records in the two catalogs are mostly independent of each other, while some features are similar. The catalog and object are linked using the MARC 856 field.

7.5 Format, Resolution and Compression of Digitized Materials

The formats for retroconverted text-based digital deliverables are:

- SGML markup;
- Database.

Some texts contained non-Latin scripts. OmniPage Pro OCR software is used to create the Library's Centennial Catalog. As this process is ongoing, no reliable accuracy figures are currently available. No special treatment is carried out prior to OCR. Keying in is used for the Library's eighteenth century manuscript full text projects. Full text materials acquired from vendors are parsed, indexed, and served through Philologic.

For image material TIFF and JPEG file formats are used for capture and preservation, while JPEG and GIF are used for delivery. PDF format is used for electronic reserve material. PSD (PhotoShop) format is used for oversize material, and the images are then stitched together. The capture and preservation resolution can be 300, 400 or 600dpi. Delivery resolution is 72, 300 or 400dpi. Capture, preservation and delivery bit depth is either 8 or 24 bit. Average file size for capture and preservation is the range of 1-80 MB. Average file size for delivery is 100KB to 5/700KB. JPEG image compression used is for delivery to improve access. The project retains the original scans in uncompressed form.

The dynamic range of equipment is checked using Kodak targets and the info tool in

PhotoShop for RGB values. The project carries out some post processing on images using Silver Fast and Epson Twain Pro for gamma adjustment and PhotoShop for de-skewing. Debabelizer Pro software is used to create access versions (GIF and JPEG).

For others starting work in digital imaging, the University of Chicago Library's digital projects would recommend ensuring that all capture devices are calibrated to the same colorspace.

RealAudio file format is used for the capture and delivery of sound material. Sound material is streamed at several different rates for different connection speeds.

The quality control procedures in place for the digital deliverables, are a total check on each image and a random check of files written to CD. There are metadata quality control procedures that ensure correct entry into the MS Access database.

Users have open access to the Library catalog and digitized materials. Searching and browsing facilities are keyword and Boolean operators. Users do not have to pay to use the digital deliverables, but instead only pay for high quality TIFF images they order.

Potential users of the digital deliverables are informed about their availability through website announcements and local press releases.

7.6 Evaluation, Funding and Long-term Sustainability

The Library has not carried out any formal front-end, formative or summative user evaluations.

The main funding sources for the various digitization projects have been the NEH, Library of Congress/Ameritech, LSTA, Save Americas Treasures and the Women's Board (of the University of Chicago Library). The Library is unsure if the use of standards has saved money but is convinced that it is the best long-term approach. Most funding organizations have monitored projects through regular reports. The Library of Congress/Ameritech were more involved and worked in detail about scheduling. Funding organizations have requested project management plans, cost models and workflow reports as parts of grants.

New material and metadata is added on an ongoing basis but no specific frequency is set. Metadata is changed as and when required and the user interface and file formats will possibly be changed. For archiving purposes the Library archives to CD and University tape backups, but there is no procedure for checking file integrity as yet. A long-term preservation strategy has not yet been written.

It is intended to keep the digital deliverables available indefinitely. The projects do not depend on self-generating funds for long-term sustainability. The Library does not have an exit strategy for its digital collections, however, it is agreed that the loss of the digital deliverables would matter.

7.7 Additional Comments

The Library believes that surveys such as this need to acquire information on the hidden costs of staffing and in-kind support that are common to the majority of institutional projects. This they feel is a critical point and that other projects or institutions looking to start digitization need that layer covered before they can begin. In addition, there is the importance of website design, while public service expertise is often needed for effective delivery of materials.

valid xhtml 1.1
abp~04/02

9 Library of Congress, Geography and Map Division

Colleen Cahill, the Digital Conversion Co-ordinator at the Geography and Map Division of the Library of Congress, was interviewed by HATII on September 19 2000. The Geography and Map Division holds the largest cartographic collection in the world, with 4.5 million maps dating from the 14th Century. The Library of Congress has an aim to make its resources available to Congress and to the American public, as well as preserving the material for future reference. To this end, the Library established the National Digital Library, whose initiatives seek to improve accessibility and encourage lifelong learning, while creating standards for a professional service in the field of archival digitization. One such example is the American Memory project, through which map collections dating from 1500-1999 were converted to digital form.

9.1 Organizational Digitization Program and Policy

A collection survey was not reported as central to the activities of the initiative. The project did not establish local digitization priorities and, while the Geography and Map Division (G&M) undertakes collection surveys, it does not establish the priorities for digitization. The policy of Special Collections of the Library of Congress (LC) has some bearing in this area, but the Geography and Map Division has the final say on what is scanned and in what order.

Obstacles to planning and building the digital deliverables, included the size of the original materials which can require up to twelve scans, moving files and stitching them together, and the use of MrSID for compression. However, the Division recently discovered a method to put the pieces back together using MrSID so that they are now only limited by the hard drive space available. With the right set up, they suggest they could put terabytes of data together.

The selection criteria and their rank for digitization are that:

- the original is out of copyright;
- digitization will improve its functionality;
- it has research significance;
- it is cataloged.

The project did feel that, within this framework, any of the prioritization criteria applied. These criteria have changed over time; in particular demand has increased for “current news” maps, which are driven by news headlines. In addition, about 12.5% of digitization is “on-demand”.

Developing the digitization program involved co-operation across archives, libraries, foundations and charities (Center for Geographic Studies), and commercial organizations (Microsoft). Based on this experience, where organizations have very different structures, objectives and needs, the program recommends that partners are flexible, so that no organization has to surrender too much.

The current status of the project is ongoing with no anticipated end date; it started in 1994 and received National Digital Library (NDL) status from 1996. The main purposes that lay behind the creation of the digital deliverables were enabling research, facilitating public access, broadening access and assisting in the preservation of the original materials. The program produced an informative statement of intent, which was explicit about its rationale and level of faithfulness to the original, through standard publicity and Cable News Network (CNN) for maps online.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- unbound printed documents;
- printed illustrations;

- hand-written documents (manuscripts, e.g. American Revolution and Civil War);
- textiles.

The nature and format of the materials digitized include velum, paper, fiber and books (some dis-bound). The largest items were 24" x 36" in size (but through stitching the original can be much larger), required twelve scans and produced a file of some 2.1 gigabytes. The project attempted to be as inclusive as possible and included Civil War maps, and state and railroad maps. The project does not intend to re-purpose the digital deliverables.

The following standards, guidelines or tools are used for representing content:

- MrSID.

The following standards, guidelines or tools are used for describing content:

- MARC.

The following standards, guidelines or tools are used for controlling data values:

- LC subject headings.

The program did consult existing guidelines for digitizing particular document types when planning its digitization strategy, including the NDL guidelines and others through its Geographic Information Systems (GIS) partners' research in such areas as resolution.

Currently HTML is used to represent structure, but the project intends to move to XML during the next two years. In relation to standards in general the project recommends that those setting out to undertake such work should:

- do plenty of research with the aim of meeting the needs of the widest possible audience;
- recognize that they will need to learn as they go;
- recognize the limits of technology and plan to respond to changes in technology, but avoid attempting to work at the cutting edge.

The intended audience for digital deliverables includes:

- K-12;
- community college;
- four-year college;
- graduate school;
- public library users;
- government;
- public sector.

The project did not feel that any one of these groups had been made a special priority, rather they tried to serve them all. Some groups may, however, have found the material difficult to use because there was little interpretative information provided beyond the catalog record. The program did not undertake an evaluation of the target audience. Groups other than the target audience could use the deliverables, but whether they had special needs is unclear. The project has taken account of the World Wide Web Consortium's (W3C) "Guidelines for Web Site Accessibility". The profile of actual users has been slightly broader than expected (e.g. K-12 access). The project did not restrict how the digital deliverables could be used.

9.2 Project Management and Planning

Advice on managing the program was available in-house. The project is considered an integral part of the Division and is supervised by the Geography and Map Division, not the NDL. The project has two co-ordinators reporting to one chief, which has resulted in tighter control. One management procedure that did not work was the original digitization workflow. The original workflow took 4 hours to scan, process, catalog and have a map ready for the internet. Through technological and other adjustments the Division have reduced this to one hour. The managerial quality assurance procedures in place are two "eyeballings" of the deliverables and two catalog reviews; these were developed from in-house experience and the project team believes that they have been successful.

A pilot study for scanning was undertaken in the Geography and Map Division, not as part of NDL. It aimed to establish training needs, technical feasibility, and workflow analysis piloting. This study led to a change in the

design of the project as equipment choice changed from Unix to NT. In addition benchmarking studies were undertaken for technology, with a base of 4-hour scans. The project formerly allocated work on the basis of one person per project, but now projects are split into sections, cutting down on confusion and lessening the burden on the person responsible for scanning. Human Resources use both job descriptions and performance indicators for all positions.

Digitization is carried out in-house rather than outsourced because the material is so fragile. Equipment was already available in-house, but NDL added post-processing equipment and one scanner. The process of flat bed scanning (using Tangent Artisan 2000) was the only option available because of the fragility of the original material. The use of a high-end camera similar to the Phase One is not viable because of cost, not technological reasons. Guidelines for data capture procedures include calibrating the Tangent scanner and handling advice from preservation staff. Color chart benchmarks are also used when a problem arises.

9.3 Human Resources and Training

People working on the project (full time equivalent (FTE)) and their capacities are:

- two co-ordinators;
- two scanners/post processing/cataloging (includes digital conversion specialist);
- one contractor for scanning and post processing;
- one technician for retrieval and preparation of documents.

The people working on the project have a variety of backgrounds — two with MLS degrees, and two others with library sciences and technical backgrounds who also have a bachelors and a masters degree. Two staff were redeployed from other areas and two were hired especially for the project. Advice on the technical aspects of digitization came from one internal and one external source.

The training needs of the project team were assessed on the job and were identified as:

- preparation and handling of materials for digitizing;
- technical operation of digitizing equipment;
- post-digitization processing (PhotoShop, MrSID & Image Alchemy).

All staff members were engaged in training. The training was organized through in-house services where possible. It took advantage of organizations' own consultants for cataloging, involved attendance at external courses for PhotoShop and XML, and learning on the job. This organization of training has met the needs of the project but a specialized class was required for PhotoShop.

9.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables. Due to the fact that the LC is a US government agency the Division can own no copyright on anything. All the Division's scanned materials are in the public domain with the exception of one map that the publisher asked the Division to scan and there is a note in the bibliographic metadata to this effect. The copyright or rights status of the final digital deliverable is not declared. Users of the digital deliverables are allowed to make printouts on paper and film, burn to CD, DVD etc., and download to a PC, LAN or WAN. Users can download and view thumbnails and lower quality images. MrSID files are 24 bit color and 300dpi when viewed in the stand-alone viewer. The TIFF file can be retrieved from the MrSID file through the stand-alone viewer. The Geography and Map Division do not permit the TIFF file to be accessed through the web, but will burn CD-ROMs via the Photoduplication Division. The highest quality TIFF images are archived. At this time no electronic management system, such as watermarking, is in use.

The project has a conservation procedure for the original materials. If materials are too fragile they are not scanned but sent to preservation. Any risk to the material identified in preparation for digitization is assessed by specialists rather than by the project. The project tries to minimize risk to the sources. Scanning is done through Mylar casings and uses cradles and supports. Some material is prepared by curatorial or preservation staff before digitization but not monitored by them during digitization (they recognize the limits of their expertise). As yet no access restrictions have been placed on material after its digitization.

The cataloging system in place prior to digitization is MARC to which the project has full access. Index and bibliographic data are used from these records in the digitization process. No further core reference material has

to be located for the digital deliverables.

Sometimes the materials have to be altered from their original format for the digitization process (e.g. dis-bound) and some material is rejected before digitization if it is too fragile. Occasionally reproductions or intermediaries are used. These take the form of paper facsimiles; the project does not like to print from microfilm/fiche.

The original material is cataloged in MARC records which includes fields for the digital images. An in-house Microsoft Access database, which is used for tracking, houses structural metadata, the bibliographic metadata is in the MARC record. Tools for controlling data values are LC subject headings. The metadata details recorded include information about the original object, the digital object, the digitization process, technical details, staffing details and administrative information. Metadata records are created primarily by the digitizer, but all other staff have some degree of input. This metadata record is then included in the main catalog, which is held in electronic form on an intranet server. The relation of the records for the digital deliverables to those for the original digitized materials is a mixture of independent and identical. The catalog and the objects are linked through a hook to the library record.

9.5 Format, Resolution and Compression of Digitized Materials

The project did not use special standards for dealing with textual materials. For images it used the TIFF file format for capturing and preserving, and GIF and MrSID formats for delivery. Capture and preservation resolution is 300dpi and delivery resolution 72dpi. 24-bit depth is used for capture and preservation and 8-bit for delivery. The compression used is MrSID, and items are compressed up to 22:1 for delivery.

The project retains images in uncompressed form and carries out post-processing stitch, rotate, crop, color balance, unmask and sometimes hue and saturation using PhotoShop. The average capture and preservation file size is 210 megabytes. Delivery files are approximately 10 megabytes for delivery.

The dynamic range of the equipment is not checked. From its experience of digitizing images the project would recommend achieving as much capture as possible at one shot.

Two people implement quality control procedures by checking each image; metadata recording is also reviewed (although the Access database is not as heavily reviewed because some material which has been scanned is not cataloged).

Users access the digital deliverables through open access to the catalog plus the materials. Users can see a prior graphical or record entry before viewing; metadata searching is handled through a link. Apart from browsing the deliverables, users are able to manipulate images by zooming in and out. Image Alchemy was used for creating thumbnails and Remedy to request server space; MrSID, Cute FTP (for bit transfer rate), Internaweb file browser and Voyager for the catalog are the specialist software tools used by the project.

The project has one of the highest, if not the highest, level of usage in the NDL program. This is monitored by information technology services to which the NDL has access. Users do not have to pay for the use of the digital deliverables. Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, print and broadcast media coverage, conferences, meetings and electronic and conventional mail shots. The project believes that electronic media (web and email) are the most effective.

9.6 Evaluation, Funding and Long-term Sustainability

No front-end evaluation has been carried out, although the NDL may have done some work here. Formative evaluation took the form of a great deal of input from staff and chief, and this led to a number of web page changes. No summative evaluations have been undertaken. The NDL is planning such evaluation, but not at the level of this project.

The project is at the high end of NDL funding. LC feels that standards have saved money by increasing productivity. The project was required to provide cost models by one of its funding organizations.

New material is digitized and added monthly (100-500 images) and metadata updated at the same time. The user interface is updated yearly and this year has new navigation tools. File formats may be changed in the future.

The project has a preservation strategy to ensure long-term access, but this is largely determined by the LC's IT Service which provides storage on the LAN. The strategy for image preservation embraces file formats and updating metadata (the project's responsibility), and storage media and storage conditions (IT Service's responsibility). Quality control procedures in the life-cycle management are producing schedules for material to archive. The digital deliverables will be available indefinitely and the project does not need to rely on self-generating funds to sustain the resource, as it believes NDL funding will become permanent. The project has an exit strategy based on the G&M Division adopting the project, but in an altered state.

9.7 Comments and Conclusions

CDP's success lies in the dedicated staff working for the project, not only in the small management office, but also in the practical areas of the projects. There is an understanding of true collaboration, and the standards and guidelines devised have ensured that all projects play an equal role. The CDP has used the project management structure to achieve usable minimum standards which projects can apply to their imaging while remaining viable in the union catalog. The DPL imaging lab, which was visited in order to see an actual project in operation, is both large and impressive with superb equipment and staff. Interestingly, they all agree that the best digitizers are photographers and not IT specialists. This is logical, for while the processes may be on computer, the knowledge of color and tone are that of a trained photographer.

The committee structure is ideally placed to represent the views of the partners and to develop the guidelines for the CDP as a whole. Training for all is an essential part of the collaborative as well as the practical side of the Project. Flexibility in the union catalog and the format of finding aids have ensured that all the projects' digital objects are available to all. There are some important lessons to be learned in collaboration from this project, perhaps the main one being that it takes a separate project management team to consolidate and co-ordinate that collaboration.

valid xhtml 1.1
abp~04/02

10 Library of Congress, Preservation Reformatting Division

HATII interviewed Lee Ellen Friedland, Senior Digital Conservation Specialist at the Preservation Reformatting Division of the Library of Congress, on September 21 2000. The particular aim of the Division's Preservation Digital Reformatting Program, is to preserve fragile or unstable materials, while also enabling them to be accessed, without further endangering their longevity. The Program therefore began the process of creating facsimiles, whether in the form of microfilm, paper copies or digital reproductions.

10.1 Organizational Digitization Program and Policy

The Preservation Reformatting Division's digitization strategy developed from a need to incorporate digital technology into the preservation reformatting program. The program tapped into the annual collection surveys undertaken by divisions throughout the Library, which nominates material for preservation. As collection surveys are carried out across the Library of Congress, the whole institution is aware of this process and staff are involved. The curatorial and collection management staff in each Division carry out each survey. A lesson learnt in this process was to issue guidelines for selection criteria.

The collection surveys are used to establish priorities for digitizing holdings. The selection criterion establishes which material is most appropriate for analog v. digital preservation. The division makes these decisions and a workplan is then devised, based on these criteria. These priorities are formalized in a strategic policy statement, which was based on that for preservation microfilm and photocopying. The Senior Digital Conversion Specialist and ultimately the Division Chief, who has to manage the budget, approve these decisions.

The objectives that this policy attempts to achieve are, firstly, the most appropriate preservation strategy for the collections and secondly, the use of digital material in the most appropriate way. In these respects the project has been successful. The recommendation from this experience is to have sound selection criteria (for the Library of Congress environment) and to ensure they are explicit but flexible. See policy documents on web: http://lcweb.loc.gov/preserv/prd/presdig/presselection.html

Overall obstacles to planning and building the digital deliverables are finite resources, the under-supply of certain technical support and a slow bureaucracy for acquiring tools.

The guiding criteria for selecting materials for digitization were primarily conservation, historical and cultural value, preservation and security, followed by improved functionality and enhanced access. The project did not feel that within this any prioritization criteria applied, nor did the criteria change over time.

Developing the digitization program took place in co-operation with other parts of the Library of Congress. Based on this experience the program recommends that work has to be collaborative, and that it is essential to have good communication and clearly articulated expectations of what other parties will do.

The current status of the project is ongoing with no anticipated end date.

The first and main purpose in creating the digital deliverables is preservation; the second, to be able to serve the digital surrogate regardless of the level of distribution (e.g. reading room only, campus-wide, web, etc.); and the third as experimentation.

The program did produce an informative statement of intent that was explicit about its rationale, long-term sustainability and level of faithfulness to the original.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- unbound printed documents;
- typescript documents;

- printed illustrations;
- hand-written documents;
- intermediaries (photographic negatives, microfilm);
- mechanical copies;
- hand copies;
- photographs.

The nature and format of the materials digitized embraces any paper or photographic based material. Everything recommended by the preservation strategy is digitized (i.e. not only a “sample”, or “entire body of material”). The program does not intend to re-purpose the digital deliverables.

The following standards, guidelines or tools are used for representing content:

- TEI;
- JPEG;
- SGML.

The following standards, guidelines or tools are used for describing content:

- MARC;
- TEI header.

The program did look to existing guidelines for digitizing particular document types when planning its digitization strategy and built on the American Memory and Making of America (MOA) models.

The following standards, guidelines or tools are used for representing structure:

- TEI;
- HTML;
- SGML.

In relation to standards in general the program would recommend using standards to serve goals (e.g. TEI and MARC are used officially in LC).

The program did not undertake an evaluation of the target audience, but other users could use the digital deliverables. The main limitation to the use of these deliverables is intellectual property rights (IPR) and this is clearly stated.

10.2 Project Management and Planning

Advice on managing the program was available in-house (via experience of senior program staff and division). The management of the program is fully integrated into the preservation services, and support is provided to all library collections. The digitization program led to the addition of another layer to the divisional level. The formal management procedures in place are those of the whole divisional level program. The managerial quality assurance procedures in place are the tracking of work by the Project Manager and Division Chief.

Both feasibility and pilot studies were undertaken on technical feasibility, workflow analysis, workflow piloting and technology forecasting to provide information to assist project management and planning. These studies led to a refining of processes in the project. In addition periodic, limited benchmarking studies were undertaken to determine the viability of scheduling tasks. The parameters taken into account in these benchmarking exercises were the technical results of convergence and quality procedures. The tools used to aid scheduling and planning include spreadsheets and databases. Who undertakes what work is determined by the outsourcing of conversion, which leaves internal staff to manage the program. Both job descriptions and performance indicators are used, the latter for contract deliverables.

Digitization is outsourced as it is not feasible to provide these services in-house. The vendor uses flatbed scanners, film scanners, high-end professional digital cameras and book cradles. Guidelines for data capture procedures are formally established in the contracting instruments.

10.3 Human Resources and Training

The people working on the project have a variety of backgrounds — library, graphic arts, photography, humanities research and project management. Staff members were not redeployed from other areas, but some microfilm technicians assist in the image quality review process. Advice on the technical aspects of digitization was predominantly available in-house.

The training needs of the project team were assessed in the applied environment of the project and were identified as project management, application of technical standards and the preparation and handling of digitizing equipment. Training was organized in-house, using the organization's own consultants, by attendance at external courses, by independent study and by learning on the job. This organization of training has met the needs of the project.

10.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables but the LC does not own the copyright to all the original materials. Every variant of copyright ownership is encountered, the predominant form being held by the original donor. Where material that is in copyright is digitized this is carried out under the legal provisions for libraries. The project declares the copyright status of the digital deliverables in a collection level statement, in accompanying documentation at the item level, and/or in bibliographic and administrative metadata or text headers. The copying rights for users vary for every set of materials. No definitive statement can be made on the format of the digital deliverables that users are able to view and download as the delivery end is the least established and is still in development. At this time no electronic management systems, such as watermarking, are in use.

The digital preservation program is fully integrated into the LC's preservation and conservation work and preservation analysis is undertaken on material as part of each division's collection survey. Any conservation work on the original materials, such as re-housing, is carried out as part of this. Whether any original material is modified or degraded to carry out digitization depends on the collection materials themselves. A risk assessment in preparing materials for digitization is carried out by conservators and an appropriate plan is created with their advice. This may involve specifying special equipment such as book cradles, special ventilation or workspace requirements. If any preparation of materials is required before digitization, preservation staff carry this out. Once the material has been digitized, access to the originals is sometimes restricted and this is determined by the curatorial and custodial authorities for that collection.

The cataloging system in place prior to digitization is the LC online catalog that incorporates a wide variety of access aids and to which the project has full access. All of the available information from these records is used in the digitization process. This is supplemented by locating further core reference material for the digital deliverables.

The materials sometimes have to be altered from their original format for the digitization process to take place, but material is only rejected if there is irreparable damage to the original. Where only reproductions or intermediaries are available the project digitizes from these. The reproductions or intermediaries used take the form of photocopies, photographic prints and microfilm.

The original material is cataloged in the LC catalog, which the collection division facilitates and co-ordinates. The digital deliverables are cataloged using MARC and USMARC standards.

10.5 Format, Resolution and Compression of Digitized Materials

The format for retroconverted text-based deliverables is SGML markup (TEI in Libraries Guidelines Level 1 for archival versions) and PDF for output only. None of these texts contained non-Latin scripts. The OCR software used to convert digital images is Prime Recognition.

TIFF file format is used for capturing, preserving and delivering images. GIF and PDF are also used for delivering images.

The project retains original scans in uncompressed form although not all master files are uncompressed. The project's vendor checks the dynamic range of the equipment and undertakes processing on images for color balance, gamma corrections, etc.

Total check and statistical sampling quality control procedures are in place, while metadata recording and

delivery are reviewed (tools are also being developed). The effect of these quality control procedures on workflow and policy decisions varies project by project.

Users do not have to pay for the use of the digital deliverables but access them through open access catalog, open access to catalog plus the materials, and restricted on-site access.

The usage of the digital materials is monitored by automatic data capture.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, conferences, meetings and electronic mail shots. The project has not yet decided which of these media is the most effective.

10.6 Evaluation, Funding and Long-term Sustainability

No front-end, formative or summative evaluations have been undertaken.

The project receives its funding annually from the LC as a percentage of an appropriated budget. In the long-term the project feels that standards have saved money.

The digital deliverables will be available indefinitely and the project does not need to rely on self-generating funds to sustain the resource. The project does not have an exit strategy.

valid xhtml 1.1
abp~04/02

11 Library of Congress, Prints and Photographs Division

Phil Michel, the Digital Conversion Specialist at the Prints and Photographs Division of the Library of Congress, was interviewed by HATII on September 19 2000. The Division holds more than 13.6 million images in its collections, prompting the creation of the Digital Library Group project, in order to develop a strategy for digitization. Preservation played a significant role in the decision to digitize the materials, while issues of public access, research and education were also considerable incentives.

11.1 Organizational Digitization Program and Policy

The Prints and Photographs Division (P&P) developed its digitization strategy from an exploration of electronic surrogates that began with an optical disc pilot project in the 1980s. The P&P Division did not conduct a collection survey as part of planning its digitization strategy.

Although a collection survey was not used to establish priorities for digitization, the American Memory project for education had overlapping interests (saving, handling, etc.) so that the collection used met multiple criteria. This established objectives for serving an educational function (K-12 and college) through digitizing popular or hard to serve material.

The project does not have any formal selection criteria or prioritization.

To date the project has not co-operated with other organizations in developing its digitization program.

The current status of the project is ongoing. The NDL pilot ends in 2000 but is ongoing with no anticipated end date. The project started in the 1980s, the American Memory pilot began in 1990, and the NDL pilot in 1995.

The purposes in creating the digital deliverables include:

- preservation;
- public access;
- teaching and learning resource;
- research, experiment;
- wider access;
- response to previous demand.

The project did not feel, however, that it could rank these aims in an order of priority. The program produced an informative statement of intent, which was explicit about its rationale.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- unbound printed documents;
- typescript documents;
- printed illustrations;
- original artworks;
- intermediaries;
- mechanical copies;
- photographs.

Across the whole project the predominant nature and format of the materials digitized consists of negatives. The digital deliverables are intended to represent the entire body of material. There is no specific intention to re-purpose digital deliverables.

The following standards, guidelines or tools are used for representing content:

- JPEG;
- SGML;
- MARC (primary).

The following standards, guidelines or tools are used for describing content:

- MARC;
- EAD (less used).

The following standards, guidelines or tools are used for controlling data values:

- LC subject headings;
- name authority files;
- Thesaurus for Graphic Materials.

The program looked at other projects as well as existing guidelines for digitizing particular document types when planning its digitization strategy.

The following standards, guidelines or tools are used for representing structure:

- SGML;
- organization of Unix file system to enable interaction between name files and MARC records.

In relation to standards in general the project recommends a system of adopting, modifying and emulating.

The intended audience for digital deliverables includes all of the teaching and learning group:

- K-12;
- community college;
- four-year college;
- graduate school;
- lifelong learning;
- distance learning;
- computer-mediated learning.

However, the project also feels it is serving other groups of the general public, subject specialists, and the government and private sector, and could not give them a prioritization ranking. The profile of actual users has been as expected.

The project undertook an evaluation of the target audiences, see: http://memory.loc.gov/ammem/usereval.html

The interviewee was unsure if the project had taken account of W3C's "Guidelines for Web Site Accessibility" on users with disabilities. As far as possible, the project determines any limitations to the use of the digital deliverables but ultimately it is the responsibility of the user. The most common form of limitation is copyright, which requires a letter of permission.

11.2 Project Management and Planning

Advice on managing the program has come from a mixture of external and in-house sources but not on a systematic basis. The management of the project is becoming more and more integrated into the P&P Division. The project has led to changes in organizational procedures as policies for serving collections have altered and new positions have been created. Less formal project management procedures are in place as part of the Division. No particular project management procedure has failed; the project has felt its way through, been flexible and altered its workflow accordingly. The quality assurance procedures have been "find and fix" and developed on an incremental, maintenance basis in the same way as traditional collection management.

Neither pilot nor feasibility studies have been undertaken. It is not known if any benchmarking studies were undertaken. Gantt charts are used to track pictorial scanning. The project timing, and hence allocation of work, is geared to the contractors' start and work is "ramped up" from there.

Approximately 90% of digitization is outsourced but an increasing amount is done in-house. Outsourcing was chosen for reasons of efficiency, expertise and the requirement of high volume in a short space of time.

11.3 Human Resources and Training

The project employs one part time (PT) director, two catalogers (PT) and five technical support staff (100%). The director and catalogers have library backgrounds, whilst most of the technical support staff have come from NDL. No staff member was redeployed from other areas. Advice on the technical aspects of digitization was a combination of external and in-house.

The training needs of the project team were assessed in an *ad hoc* fashion and were identified as the technical aspects of digitization and post-digitizing processes. All team members engaged in training and this was organized through in-house consultants, external consultants, external courses, independent study and learning on the job.

11.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables but does not own the copyright of the original materials. The copyright or rights status of the final digital deliverable is declared in a copyright statement at the collection level, and this has been effective. The material digitized in copyright was done with the owners' agreement.

Users of the digital deliverables are allowed to make printouts on paper and film, burn to CD, DVD etc., and download to a PC, LAN or WAN. For some of the NDL collections, users can download and view thumbnails, lower quality images and highest quality (up to TIFF) images. For the P&P collection users can view and download thumbnails but lower and higher quality images are only available internally for some of the collections (on-site). No electronic management systems, such as watermarking, are in use.

The project has a conservation procedure for the original materials (part of P&P duties) and material is investigated prior to digitization. If necessary, conservation staff undertake conservation activities, and no materials are modified, degraded or compromised to carry out digitization. Risk prevention methods for the material in preparation for digitization include safe handling practice from conservation to the contractor and close attention to the work environment. Steps to minimize risks are built into the contractor plan. The materials are stabilized by curatorial or preservation staffs prior to digitization but not monitored by them during digitization. After digitization, access to some of the original materials is restricated; the digital surrogate is seen as the primary source for users.

Some cataloging systems are in place prior to digitization (such as lists or inventories) and any information available from these is used in the digitization process. Although the project does not have access to all the relevant cataloging information, staff know at least the quantities and material types, and the digitization process sometimes feeds back into the catalog record — e.g. negative size. Some core source material may be located but this is unrelated to the digitization process. The only way in which material may be altered for digitization is sometimes to lift sleeves, Mylar or tissue. No material is rejected prior to digitization. The project sometimes uses intermediaries, where these already exist (such as transparencies of posters, slides, 35mm, 4x5 transparencies, photographic prints).

The original material is cataloged in MARC and EAD records. The digital surrogates are referenced by bibliographic object using MARC, EAD and USMARC standards. Tools for controlling data values are controlled vocabulary (MS name rules), thesaurus (Thesaurus for Graphical Material), and a classification scheme. The metadata details recorded include the original object, the digital object, the digitization process, technical details and administrative information.

All members of staff create different parts of the metadata records. The catalog for the digital deliverables is held in paper and electronic form on an intranet server and technical details about the project are also available on the web: http://www.memory.loc.gov/ammem/techdocs/white.html.

11.5 Format, Resolution and Compression of Digitized Materials

For images, TIFF file format is used for capturing, preserving and delivery, while GIF and JPEG are additional delivery formats. Capture, preservation and delivery resolutions range from 50-3500dpi. The original bit depth used for capture, preservation and delivery was 8 but is now 12 to 16. JPEG compression is used for delivery

only. TIFF Gp IV compression is used for best copies of texts and architectural drawings (bitonal images).

The project retains images in uncompressed form and only holds compressed bi-tonal images. Post processing operations are limited to skew and gamma adjustment on negatives as the project has a strong emphasis on full capture across the range of material right up to 200MB master files. The dynamic range of scanning equipment is checked by the vendor and was a factor in selecting it. From its experience of digitizing images the project would recommend benchmarking, doing tests, taking time at startup and not taking specifications at face value.

The quality control procedure in place for the digital deliverables takes the form of a quality assurance plan from the vendor (who carries out total checks). The project also does spot checks on the TIFF headers. The vendor also has a few systems for metadata quality control, but these were not specified *per se* by the project. Quality control procedures have had a direct effect on the management of production workflow. There are now more direct look-in sessions and smaller batches to the vendor.

Users do not have to pay for the use of the digital deliverables and instead have open access to catalog plus the materials. Standard web browsing is offered which also links to the catalog. The system does allow searching of metadata but in a limited way. Special software used by the project includes inquiry software and a relevance ranking search engine. These required extra programming on top to control the display.

Usage levels are monitored by information technology services (ITS) to which the NDL has access.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, print and broadcast media coverage, conferences, meetings and electronic and conventional mail shots. The project does not know which medium has been the most effective.

11.6 Evaluation, Funding and Long-term Sustainability

No front-end or formative evaluation has been undertaken. A summative evaluation has been undertaken.

The project is not aware of how much it has cost to date. The main source of funding has been the NDL, which has received gift funds from various donors, with a little from the LC. In relation to how much the project should have cost, the project feels it has been the right size, pace and scale for its funds. The project was less sophisticated five years ago; if it had its funding over again it would have higher productivity now. The project believes that standards have saved money. It does not know how the funding organization monitors the project or if it asked the project to provide collections surveys, cost models, project management plans, etc.

Under the NDL, new material (and its associated metadata) is updated two or three times a month. Metadata are changed weekly and the user interface is not updated systematically.

The project has a preservation strategy under development based on file format, storage media, storage conditions and updating metadata. Quality control procedures are in place for life-cycle management but the details are not known.

It is intended for the digital deliverables to be available indefinitely, since the project does not rely on self-generated funds in order to sustain the resource. The project does not have an exit strategy.

12 Library of Congress / National Digital Library Program (NDLP)

On January 22, 23 and 24 2001, HATII interviewed the following members of the National Digital Library Program at the Library of Congress: Martha Anderson, Project Manager; Deborah Thomas, Digital Conversion Specialist; Tamara Swora-Gobel, Digital Conversion Projects Co-ordinator; Karen Lund, Digital Conversion Specialist; Marc Dudley, Computer Specialist; Thomas Bramel and Mary Ambrosio, Systems Analysts; and Dave Woodward, Computer Specialist. The core objective of the NDLP was to digitize the five million objects within the Library's collection of Americana, which was successfully achieved in September 2000. The American Memory project served as a pilot study, leading to improvements and an increased emphasis on standardization for future initiatives.

12.1 Organizational Digitization Program and Policy

The National Digital Library Program (NDLP) is a flexible newer unit of the Library of Congress that was set up specifically for the creation of this program over a five-year period (1996-2000) with the financial support of the Congress and private funding. In September 2000 it achieved its target (set in a rather arbitrary fashion at the program's initiation by the Librarian and the political/funding influences without real input from the NDLP team) of digitizing 5 million objects. The program is now moving towards a second phase when a lot of discussions are taking place about its future, integration with the other existing departments of the LC, funding and *modus operandi*. It is anticipated that the new phase will be an ongoing process and service, without an end date.

12.1.1 Collection Survey

Neither the LC nor the NDLP has a formal digitization strategy as such, but several activities led to the formulation of clearer objectives and digitization working practices. They carried out an optical disc pilot in the early 1990s that sought to capture the primary photographic collection in digital format. Then in 1995-6 there were discussions about the distribution of digital material. They began with what was possible for technical reasons (e.g. photographs), what was free from copyright, without any legal restrictions, and depending on technical feasibility, were influenced by the Librarian's and the custodians' desire to provide access to unique materials. There was some consideration of what was heavily used in analog form, but this was not the primary motivation.

A collection survey was carried out, in the process of planning the digitization program on American history and culture. The ten custodial divisions were interviewed and 200 collections were selected. The type of questions asked in the survey were:

- What do you think is technically feasible?
- What would be interesting?
- Is the material in a condition / easy to be scanned?
- Is it copyright-free?

They also looked at educational recommendations, and material that was in high demand by researchers and the public. The institution as a whole was made aware that the collection survey was being conducted and staff were encouraged to participate. However, it seems that it was left to the individual custodial divisions to organize who would participate and the nature of their contribution. Three or four people in the NDL Task Force team carried out the collection survey. The survey produced good results, providing a body of work that the team could build on. There was some conflict with the custodial divisions, which were hoping that they would have more control of the prioritization and selection of the material. In some cases they thought that it would be a more scholarly exercise. Also, private funders sometimes pushed collections up the list because of their own agenda. There is more information available now about what is technically feasible and how material could be used, than was available when they started.

12.1.2 Priorities for Digitizing

The collection survey was used to establish priorities for digitizing holdings. As was indicated, these were established by the custodial divisions in conjunction with the NDL staff, with some influence from funders and their interests and priorities, and from practical aspects, such as the scheduling required by conservation. The NDL staff in conjunction with the custodial divisions would put together suggestions that were approved by the rest of the team, for example in their two- and three-year plans. These priorities were not formalized in a strategic policy statement, since the documents prepared by the NDLP were more production-oriented. There was goal-setting every year, rather than an overall strategic policy statement. It seems that the high stakes and the pressure to achieve the targets specified the documents required to facilitate production and project management, rather than more general policy documents.

The objectives of the program were primarily to produce a large database of Americana, but also to provide access to the material (not preservation), and finally to help preserve the digital product. The program has been successful in achieving all of these.

The NDL project manager would recommend to other organizations attempting to formalize their selection criteria in strategic policy statements and embarking in digitization projects in general, having a clear goal of the volume of material and work from the beginning. In this case, the fact that they knew how much material they had to digitize and within what time frame, helped to motivate the team, cut short unnecessarily lengthy discussion, moved the whole team in one direction, and helped them to focus on content. Other projects often fail when they try to do too much. It is important to agree about a program with a specific goal, rather than leaving it to discussion.

12.1.3 Overall Obstacles

In planning the development of digital deliverables, it was difficult getting to the specific details from the start. Another obstacle was developing a unified perspective for how the work should progress. With a project of this size and type, accommodating diversity of media types was also difficult. For the motion pictures, the limited amount that existed in the public domain also proved problematic. For folk tunes, the nature of the material, the fact that there are often more than one titles for one tune and several performances posed difficulties in planning.

In the process of actually building digital deliverables from the collections, the main problem was dealing with the technology (e.g. capture devices), getting enough storage, and trying to forecast accurately. Another challenge was blending the different collections into a coherent whole. The level of cataloging and the filenaming used by the various departments and staff varied, so it was often difficult for the technical team to link the descriptive records with the digital files. For example, the archivists related more to the physical location of the material, rather than use naming conventions that were consistent or appropriate to the digital world. Finally, finding physical space to accommodate the new workforce was problematic.

12.1.4 Selection and Prioritization of Materials

In both the selection and prioritization of materials for digitization the main criteria in order of importance were:

- historical and cultural value;
- enhanced access;
- intellectual property rights;
- potential to reach disadvantaged groups;
- conservation (e.g. decreased wear and tear);
- social inclusion;
- publicity.

These did not really change over time, apart from a few cases where there was pressure from outside sources, e.g. funders. So, in some cases, although these were the criteria, the practice was different. In the case of conservation, they generally tried to ensure that the treatment of the material that would be required would not be too costly or time-consuming to impede progress. They tried to select collections that would step through the workflow efficiently, since they had a target amount to reach. With the curatorial divisions more involved in the digital program in the future, it is likely that research significance will acquire more importance as a selection and prioritization criterion. Staff in large research libraries usually have their favorite collections which tend to surface as selections even when they are not the best or easiest to be converted.

12.1.5 Co-operation

Although the main activities related to the program were carried out by the in-house NDL team, there was wide-ranging co-operation with other organizations, including archives, other libraries, museums, academic

institutions, corporations (for funding), foundations and charities (as donors), government agencies, and historical societies. These spread geographically across the whole range from local to international.

This experience and the sharing of information and data with other institutions reinforced their belief in the general rule of interoperability. Another important rule was to try not to bring extra work to the house, to assess where the work should be done and by whom. Another recommendation would be to be careful about managing the expectation of donors.

12.1.6 Purpose of the Program

The main purpose for the creation of the digital deliverables was public access, provision of a teaching and learning resource, preservation, and research. The program produced a public affairs announcement and also included information on its website that made explicit the rationale, the scope, and the primary audience of the program. There is also a mission statement on every collections page on the website.

Revenue generation was never intended, except to intrigue donors and raise consciousness. Other areas of the LC, e.g. duplication services, will probably generate revenue in the future and respond to user demand. In the future they might digitize new material in answer to customer demand. They will need to explore these issues in the future – how this could be done, by whom, and whether any fees would be incurred or not. This is all presently in a state of transition.

12.1.7 Nature of the Source Materials and the Impact of Digital Deliverables

The program digitized a very wide range of source material with varied format and nature:

- printed books, serials, and similar bound materials (published);
- printed documents (gray literature);
- unbound printed documents;
- typescript documents;
- printed illustrations;
- handwritten documents;
- original artworks;
- intermediaries (film, microfilm, microfiche, photocopies etc.);
- mechanical copies (photostats);
- hand copies;
- photographic negatives (but not intermediaries);
- photographic prints (but not intermediaries);
- film;
- sound recordings (variety of forms: e.g. acetate discs, real tape, cylinders);
- video recordings;
- radio broadcasts;
- maps and cartographic material.

The materials selected for digitization were both the entire body of some collections, as well as a representative sampling. They initially intended to do much more of the former, but in practice ended up doing samplings more frequently.

12.1.8 Interoperability

For representing content they used TEI, JPEG, MPEG, and SGML. They are currently exploring the use of XML, but have not applied it yet. Only one Ameritech collection came with XML markup. For describing content they used MARC, Dublin Core, EAD, and TEI Header. Controlling data values was the area that presented most difficulties. Although they used Library of Congress subject headings, Library of Congress subject thesaurus, genre headings for films, name authority files and the Thesaurus of Graphical Material, none of these were enforced. There was not as much consistency in this area as they would have preferred. The differences in the descriptive practices among the contributors proved to be a significant and common problem. For moving images, they tried to conform with the film division of the LC and what the Library as a whole used, but it was very difficult to catalog film using MARC. Nevertheless, it is important to keep some common guidelines. Whenever the cataloging deviated from MARC, it was more difficult for the technical staff to integrate in the system.

For a public institution that tries not to incur any costs to the end user and has a goal of reaching as many

people as possible, it is even more important to stay away from proprietary formats and use commonly agreed standards. It is important to try to provide easy access to the user. Also, it is best to try to set standards at the beginning, so there is no need to go back and fix things retrospectively.

12.1.9 Target Audience

The primary intended audience for the digital deliverables was K-12. Although the program was also relevant to and could be used by community and four-year colleges, graduate schools, lifelong, distance, and computer-mediated learning, public library and archive users, these were all secondary audiences.

There was no evaluation of the targeted audience, although the existing experience of the LC and feedback from User Services was very useful. The Learning and Visitors Center gives them a sense of who is looking at the digitized materials. They also get useful feedback at conferences.

They tried to acknowledge the needs of those with disabilities. For example, in the presentation and design of the user interface they did not use frames. It was not always easy to accommodate disabled audiences in relation to content, e.g. silent films.

Where there were limitations as to use of the digital deliverables, the reasons were clearly stated on the website (e.g. intellectual rights, material only for private use).

The profile of actual users was the same as the one anticipated, although there is more general public use than educational. There are also many foreign users – they reach a truly worldwide audience.

12.2 Project Management and Planning

Project management is conducted in-house. There is a Management Team and regular meetings of the project leaders. The project director is an accountant who was very successful in raising private funding and was not closely involved in the digitization process and production.

The NDL team is very different from other departments in the LC. Although the normal federal regulations apply, the program was “fast-tracked” and there is a certain distance from the other departments that will probably change in the future. Their structure is much more even and flexible - less hierarchical. They are very team and project-based and less bureaucratic. Some of the teams that work around specific tasks are very small. They have to work with the web design team, the programs division people, and often fit with other people’s schedules. Unfortunately, although they would have liked to be able to report that the NDLP led to changes or re-evaluation of organizational relationships and procedures in the rest of the LC, this has not really happened. This is currently an issue, with all the discussions about the next phase of the program.

The initially tried to create a central Technical Liaison Production Team to provide support to all content teams, but it seems that content teams actually preferred to do this themselves, and so the idea was abandoned.

In order to ensure quality assurance, they developed guidelines for work with vendors, an editorial handbook about the quality of documents on the web, and have put a great deal of material on the intranet with resources, examples, and models. This has proved valuable and grew out of the need to make resources available.

They did not carry out a feasibility or pilot study for project management purposes. The American Memory pilot provided the core framework, but was not conceived as a project management and planning study. That pilot did lead to changes, after observing the effects on the workflow. There was increased emphasis on standardizing the relationship with the vendors. For example, they developed a database about what was received from vendors and when it was delivered. Common quality document guidelines were also developed. The American Memory Pilot included three films and was useful in order to see what could be done. The videotape was sent to 3-4 different contractors and helped them to assess turnaround time and quality. Once they decided what worked best, they stuck to it. No specific time and motion or other benchmarking study was carried out at the start of the program. They developed the project team organization based on custodial divisions. (Later they developed a core support activity team.) This was the result of experience, wanting to make progress on numbers. Job descriptions were initially very general, but have become more specific, although a great degree of flexibility is still required. Performance indicators were not produced, as the structure of the LC does not encourage their use.

Digitization was carried out in-house, however, a greater amount was outsourced. The rapid change of technology led to a preference for outsourcing, but the maps are digitized in-house as the equipment was

donated. They felt that it was not in their interests and goals to learn about the technology, so they outsourced wherever they could. Generally, the technology available outside is more cutting-edge and more standardized. However, for some materials such as audio digitizing, in-house provided better quality control. For much of the folk audio material it was important to provide a faithful digital representation of the original archival material, whereas some professional houses often "clean" the sound, deleting for example the sound of the dog barking in the background of the singer's performance. Also, private companies use various staff members to process material, so sometimes the initial instructions are lost from one operator to the other. For these reasons, it was often more efficient to digitize in-house rather than invest time and resources reviewing the material that was outsourced.

Technologies Used for Image Digitization	***Manufacturers***
Flatbed scanners	Agfa, Tangent, UMAX, Pollenex, Kurtzweil
Film scanners	SunRise
Digital cameras	JJT
High-end professional cameras	Phase One
Other	Minolta Book-key Open-book scanner, satellite scanner

Guidelines for data capture procedures (e.g. calibration of equipment, handling dust, etc.) are part of the request for proposal (RFP) to vendors. They use grayscales and color charts as benchmarks.

12.3 Human Resources and Training

The number of people who worked on the program (not counting outside contractors) were:

Type of Staff	Number	% of time on the project
Director	1	100
Metadata specialist (1 per team)	10	50

Curator	10-20	5-10
Digitizer – in maps only	4	50
Photographer	-	
Technical support staff	1	100
Technical development staff	2	75
Education specialist	10-11	100
Evaluation specialist – quality review	40-50	25-50
Digital Conversion Specialists (image management and transformation, collation, preparation of material)	40-50 40-50	10 25-50

The background and profile of most people on the team was that of the “technical humanist” – usually a degree and good background in humanities with high level of IT literacy. The team was comprised of approximately 25% trained librarians, 25% specialists (e.g. in folk life), 25% historians, and for 25% this was their first professional job. A generalist/specialist is needed for these projects. The post requires flexibility and ability to see the big picture, knowledge of the cultural picture, with ability to focus on technical issues. The work involves a combination of mundane and creative activities. Work with interesting collections is important for job satisfaction. They have much more flexible procedures than the rest of the LC, and team members try out different tasks. Very few staff were re-deployed from other areas. The NDL team is seen as very talented, committed, technically aware and responsive.

Although advice was available in-house about technical aspects of digitization, they also used external advice. The training needs of the project team were assessed on the fly. Areas where training was needed were project management, application of technical standards, preparation and handling of materials for digitization, and post digitization processes such as editing. Until now training on preparation and handling of materials was carried out on the job, but they are about to embark on a more organized way of doing this. Training was received by curatorial staff, specialist technical staff and digital conversion specialists. This was organized in a variety of ways, in-house, using project staff, the Library’s own consultants, and external consultants, by attending external courses, with independent study, and a great deal by learning on the job.

The general approach is to try to identify the core basic level of training needed and offer it to everyone, then to encourage staff to fill extra needs as necessary. It is important to hire highly skilled people with curiosity and initiative to learn in the first place.

12.4 Project Life-Cycle Processes and Procedures

12.4.1 Reproduction and Copyright

The organization is aware of the copyright position or other rights status of the digital deliverables that the program has created. They do not own the copyright to the original materials. There is a huge variety in copyright status in American Memory materials. Writers, photographers, publishers, companies, and individuals are represented. Many unidentifiable creators such as authors of older, unpublished letters are represented. The vast majority of the materials are, however, in the public domain as works of the US government or because of their age and/or publication history. As a policy matter the organization has not declared the copyright or other rights status of the final digital deliverables, particularly where the underlying work is in the public domain. As a principle, they are trying to maximize access wherever possible. They certainly recognize the considerable skill and professional judgment required to make high quality scanned images but do not treat those “plain” scans as new copyrightable works. As an aside, if the underlying work were subject to copyright, the scan would be a reproduction or a derivative work and would require the permission of the original copyright owner to exploit in any manner beyond their Library functions (i.e. for revenue generation).

When material in copyright was digitized, this was done in a variety of ways:

- under legal provisions for libraries;
- with the owner’s agreement;
- by payment to the owner of a fee;
- under license.

Copyright issues are more complicated when dealing with folk material. Although there are no performance rights in federal copyright law, the Folk Center of the LC believed that they had to try to find the families of the performers.

Users of the digital deliverables can copy the material in multiple ways at their own terminals as long as they have the equipment and facilities, since there is no control over that. Users must make their own assessment of the legality of particular works in the context of their intended use. The Program provides no warranty that everything on the site is public domain or “rights free” – it could not make nearly the volume of material available if that threshold was used. Users coming to the American Memory site are still in a “Library space” and must act accordingly. (See the Copyright and Other Restrictions Notice on the website: http://memory.loc.gov/ammem/copyrit2.html). For example, special exemptions in the Copyright Act for libraries on which they rely may not extend to a third party’s use. Users must make that evaluation for themselves. The NDLP team provides as much information as possible where rights or permissions are obtained. As a practical matter, they strategically sought for American Memory to maximize the amount of public domain content to facilitate the broadest possible use. Other projects may not have the same focus. As they begin to grapple with preservation and access for “born digital” materials and other copyrighted content, they will probably consider using copyright management systems to assure copyright owners of maximal security. They strive to balance this with appropriate access over time. They also need to be able to “collect” materials subject to other electronic management systems, which is another layer on the access, preservation, migration chain of issues.

Users of American Memory can download a variety of digital deliverables. For texts, there are ASCII text files and TEI DTD marked-up text, while they are considering XML for the future. For digital images, users can download thumbnails, lower and highest quality images and associated documents. For digital audio, users can download samples of less than 30 seconds, full length compressed and high fidelity sound files, as well as associated documents. For digital moving images, there are samples, lower (QuickTime) and highest quality (MPEG) digital video clips, RealMedia streaming files, as well as associated documents. No electronic management systems are used to control copying. They have experimented with MrSid format, but have not used it extensively. In a few cases they do not provide the TIFF image to control copying or use the user’s IP address to restrict access to some prints and photographs to LC machines.

12.4.2 Preservation/Conservation

The NDLP team worked with Conservation for assessment of how handling would affect materials. The conservationists assessed the material and took steps to minimize risk. They were involved in all the RFPs and contracts, provided tips for handling during scanning, such as on the use of cradles, and gave specific instructions per type of material or collection. They were also consulted on the space and the environment, equipment and hardware. (More information can be found in “Conservation Implications of Digitization Projects”, a paper prepared by the Library of Congress National Digital Library Program and the Conservation Division in 1999 - http://memory.loc.gov/ammem/techdocs/conserv83199a.html.)

Images of original paper prints had to be transferred to 35mm film in the lab prior to digitization, although this was done more to improve access rather than for conservation reasons. Some strengthening of material was required prior to scanning and some after scanning, in the case of disbound materials. One of the most common

dangers identified during the preparation or digitization process was damage to the binding and potential chipping. Very few bound volumes could be inverted and required face-up scanning. They used the Minolta camera for this, but lacked the option for scanning in grayscale. As a result a long queue has been created and the printed ephemera, for example, are still not completed. The materials were prepared by curatorial staff before digitization. There were always members of staff with the material at the time of scanning also for security reasons, which led to significant staff drain. They have no plans to restrict access to originals once the material has been digitized, although the Prints and Photographs division might consider it for some material.

12.4.3 Preparation and Sources

The cataloging and reference systems used at the Library are the LC catalog, MARC, and finding aids. Where available, all relevant information from these systems was used in digitization. For most materials, however, there was not enough information available (e.g. for manuscripts). The NDL team had access to all the relevant materials (e.g. cataloging information) for the digital deliverables collected before the digitization process, but had to locate core reference or source materials where information was missing. This required a great deal of work - about 60% of the effort. For example, for folk audio material information was often retrieved from finding aids, lists, data on record jackets, cards written by volunteers, and field notes.

Some of the material had to be altered from the original form for the digitization process, e.g. disbinding, removal from frames, but not much. They tried to co-ordinate with conservation and digitize material that was going to be disbound or altered anyway. In a few cases they rejected material before digital imaging, mainly due to the physical circumstance of microfilm (e.g. major variation of density in the same micro-image, micro-images of inappropriate density) and conservation reasons, e.g. brittle books. They were surprised by how many of the problems the vendors were able to overcome. They digitized originals, as well as using reproductions and intermediaries. For images they used photocopies, 35 mm slides or 4 * 5 transparencies, photographic prints and microfilm. Most audio is 1-6 minutes long. The form of intermediaries used for audio included variable speed discs, audio cassettes and wax cylinders that were converted to DAT tapes and ¾ inch U-matic videotapes. The intermediary used for moving images was Betacam SP videotape. From 35 mm film they would transfer to 16 mm film and then copy to Betacam SP tape. For some films, errors in the printing process were corrected (e.g. double imaging, upside down images that were not faithful to the original). This was documented in the “About the Collection” page.

12.4.4 Metadata

Each custodial division is responsible for cataloging the original (physical) materials using MARC in most cases. For digital surrogates, the NDL staff add the digital ID on the MARC record or create an Access database record for the manuscripts. They also used Dublin Core (Ameritech), EAD, and TEI Header for cataloging the digital deliverables. For controlling data values they used controlled vocabularies, thesauri, and classification schemes, such as the Thesaurus of Graphical Material (TGM), the LC Subject Headings, and Name Authorities. The metadata recorded information about the original object, the digital object, and they are currently in the process of adding information about the digitization process. Until now, information about the digitization process can be found on the website, but not in the individual records. In some cases, they also record staffing details. The metadata records were created by a range of staff, including digitizers, archivists/information professionals, digitizers who were also archivists/information professionals, or the digital conversion specialists. The metadata records for the digital deliverables are kept in a separate catalog, available in paper form and in electronic form on an Intranet server. The records for digital deliverables and the original digitized materials were in some cases the same, while in others independent of each other. The catalog and the objects are linked through the A56 field.

12.4.5 Format, Resolution and Compression of Digitized Materials

Technical notes available by type of material can be found at http://memory.loc.gov/ammem/dli2/

12.4.6 Text

The formats chosen for retroconverted text-based digital deliverables were HTML and SGML markup. They use OmniPack to convert SGML to HTML automatically. They have just started using OCR software (TextBridge) to convert the digital images. This was used for slave narratives with dialect, so the level of accuracy was low, at about 75%. It was still sufficient for their purpose, however, which was to be able to retrieve names of people and places through enhanced searching. This was achieved in a much more cost-effective way than keying in. Their experience shows that you need to be clear what you want from the documents. Re-keying in the case of the slave narratives would have been much more expensive with the same search result. They have also used “keying in” which is expensive. It showed a need to simplify the DTD and reduce its requirements.

12.4.7 Images

The file formats used for images were GIF for delivering, TIFF for capturing, preserving and delivering, and JPEG for capturing and delivering. The resolution used is 300-400dpi for capturing and preserving and 72 or 300 for delivering images. They used 1 to 36 bit-depth for capturing and preserving, and 1 to 24 for delivering. JPEG compression was used at capturing and delivering stage while LZW was used by vendors. The aim of the compression was to improve access. The original scans are retained in uncompressed form. Some post-scanning processing is carried out on the images, mainly scaling and producing deliverables, for which they use Image Alchemy and Photoshop. The average file sizes created were 50-300 MB at all stages - capturing, preserving, and delivery. They did not check the dynamic range of the scanning equipment in-house, but the vendors were expected to do so using targets.

Their experience showed that you should try to capture at the highest quality you can afford. It is best to let the vendor do the job and judge only on the output. It is best to specify what you want and let them decide how to produce it – bear in mind that it is not photocopying and the rate of progress may be slow.

12.4.8 Sound

For the digitization of sound materials they usually convert from the original format to DAT tape. They make two DAT tapes and one ¾ inch U-matic videotape for safety. This process was initially outsourced, but is now carried out in-house. WAVE (.wav) files are derived by sampling from the DAT at 16 bits, 22,500 times per second (22.05 kHz and 44.1 kHz sampling rate used). The WAVE files are then converted in batches to RealAudio and MP3 file formats (MPEG 2 Layer 3 that is a lossy compression) for delivery. The aim of the compression was to improve access and enhance usability. Most audio files vary in length between 1 and 6 minutes. The file sizes created were approximately 2.6 MB per minute. In some cases post-capture editing was carried out in order to remove obvious pops and clicks.

12.4.9 Moving Image

The work methods used to digitize moving images is as follows. Films are copied to Betacam SP videotape at a film production house, which creates the digital tags that are then sent to vendors. The vendors then deliver files on CD-ROM. The NDLP team takes these and creates catalog records on which the web presentation is based. They experimented with AVI in the beginning, but did not use it. They use MPEG (version 1), QuickTime, and RealVideo for capture and delivery. The MPEG version is created at 30 frames per second and at a spatial resolution (size) of 320x240 pixels at a data rate of approximately 1.2 Mbits per second of playing time. The QuickTime version is created at 10-15 frames per second and a spatial resolution (size) of 160x120 pixels at a data rate of approximately 640 Kbits per second, usually quoted as 80 Kbytes/sec of playing time. They have now started also delivering a streaming RealVideo version. QuickTime is used with the Cinepak compression algorithm, with some files using the Sorenson compression. The reasons for using the compression were, in order of priority, to improve access, to enhance usability, to decrease storage requirements, and finally to reduce cost. The average file sizes created were about 10 MB/min for the MPEGs, and 5 MB or less for QuickTime. The post-capture processing carried out on the moving images was mainly the conversion from QuickTime to RealVideo.

12.4.10 Quality Control

They used a variety of procedures for ensuring the quality of the digital deliverables: spot checks, random set of checks, check on a stratified random sample, percentage checks on carriers (e.g. corrupted disks), total check. For film material, they review all the files created by vendors, which is a very time consuming, but nonetheless vital step in the process. In a few instances they had to return material to the vendor, since the quality standards they demand are very high. The text presented about the films is approved by the film division and the NDL project manager. Recently they have hired specialized editors on the content, which has led to a more consistent and formalized presentation style. They have a receipt and review unit that runs all CD-Rs through diagnostic software to make sure that certain parameters are kept. It is very important to keep statistics on the material received, information on vendors and the quality produced. They have a three-week limit for quality review or return of batch for paper or microfilm.

There is no formal way of checking quality of metadata recording; the best way is to see whether it works or not. In general the control of metadata varied more than the checkup of the actual digital files. There was a lot of trial and error. Some of the automated procedures that the technical team has put into place pick up mistakes from the other departments. For audio material, subject specialists review all the records before they are released for the public. This process depends on scholars and their own timetables. Generally speaking, they found that for every minute of audio, ten minutes' work was required after the digitization was completed to find the file and

review it. Quality control procedures affected workflow and policy decisions, which involved coaching new staff, revisiting collections, establishing policy on handling vendors and contractual obligations.

12.4.11 Delivery, Access and Use of Digital Deliverables

Users have free open access to both the catalog and the digital deliverables for browsing, but not for manipulating in any way. The level of use of the material is very high and there is a steady increase every year. This is monitored via automatic data capture (http://www.loc.gov/stats/), although there is not much analysis. They record the browsers and platforms used. This is generally motivated by the Public Affairs Department. They also respond to queries and problem reports.

The system supports subject searching by keyword, also browsing by title, author, music performer, film title, genre, venue of jazz clubs, list of tribes. When new data is entered, these lists must be re-run. Some browse lists are actually search results, but look like browsers. Metadata fields are fed into the index, not just the subject field. Some collections allow search in all the core fields, while some others support full free text searching of e-texts. Some researchers have asked for full-text searching throughout the whole material, across collections. Although this is possible, it is not practical. There was experimentation with the Learning Center for tools to allow users to manipulate the collections, for example to create lesson plans or with page viewers for young children, but this has not been followed up.

12.4.12 Dissemination and Publicity

The methods used for informing potential users about the digital deliverables created were: announcement on website; press release; articles in print media; some print and broadcast media coverage; announcements at conferences and meetings; announcement at electronic listservs. There is no systematic way of evaluating the effectiveness of the various dissemination and publicity strategies.

12.5 Evaluation, Funding and Long-term Sustainability

Until now, very little evaluation has been done, since the emphasis was on production and creation of content.

An evaluation survey on the use of American Memory during the period 1991-93 was carried out in 1993. This was published on the web (ftp://rs7.loc.gov/pub/american.memory/user.eval/intro.txt). It included more than 40 libraries in institutions such as schools, colleges, universities, special, public, and state libraries and used 1,800 user questionnaires, 120 user interviews, and site visits. That survey clinched the argument about addressing a K-12 audience. An educational contractor was hired for Children in Technology to assess the subject areas, the content to be used, and how these related to the Curriculum, but this did not lead to any changes. Although an online study was carried out several years ago, there were methodological limitations, which resulted in a demographically limited return. LC is just embarking on a general survey that will include the NDLP.

Generally, informal feedback from the educational and user services was useful. For example, feedback from teachers about the platform most commonly used in schools led to the use of QuickTime for Macintosh computers. They also offered feedback about the presentation of material, as well as about broad themes that they would find useful. The moving images team also noted, that whenever they offered more contextual information about the collections, they always received very positive feedback. Subsequently, whenever they can, they add some contextual material, such as essays, interviews with Edison, articles from journals of the time, as sometimes “replicating the Library experience” and letting users make sense of it themselves can prove shortsighted. Even though they did not embark on extensive formal evaluation programs due to production pressures, it is always interesting for the NDLP staff to get feedback from the public, otherwise they feel that they operate in a kind of vacuum.

Informal observation and queries from the public indicate the high level of use, the educational impact, the varied audience and the effect of the resource. For example, the BBC made a program on “Voices from the Dust Bowl”, based on the American Memory material from the migrant worker collection, documenting the everyday life of residents of Farm Security Administration (FSA) migrant work camps in central California in 1940 and 1941. The Omaha Indians use the resources to learn the language at their schools (http://memory.loc.gov/ammem/omhhtml/omhhome.html). It has also received a lot of web awards, such as the Encyclopaedia Britannica’s Best of the Web and the Top 100 of 1999 of the Education Source http://www.loc.gov/loc/awards.html.

The program cost $60 million in its first five-year phase (1996-2000). $15 million came from the appropriation of Congress funds, while the remaining $45 million was raised by the LC from private donations (*Founding*

Sponsors with contributions of over $5 million; Mr John W. Kluge; The David and Lucile Packard Foundation; about 20 *Charter Sponsors* who contributed at least $1 million, such as AT&T, Ameritech, Bell Atlantic, Eastman Kodak, Reuters; about 13 smaller contributors; and three contributors in kind, such as IBM and Hewlett-Packard). The hidden costs were greater than that. The monitoring of the program from the funding organizations was fairly informal. They produced some cost models, Gantt charts and reports, but not much more. The reputation of the organization, the high profile of the program, and the support from Congress probably played an important role here.

The NDLP team feels that it was generally sufficiently equipped. They also think that using standards saved them money.

Most elements of the project will need to be updated fairly frequently. New materials and metadata will need to be digitized approximately every three months or quarterly (probably at a lower rate of production than was carried out until now). Metadata will need to be changed at a fairly low frequency. The user interface will need to be changed about once a year. It will be necessary to change file formats, but the frequency for this is unknown at the moment. There is relative safety in basing choices on a recognized standard. For example, although MPEG 4 seems to be growing in popularity, users can still use MPEG 1. Adding a new format does not necessarily render the previous one obsolete.

They are currently in the process of formalizing a preservation strategy to ensure long-term access to documents and objects in digital form and are working on their Digital Repository. This involves the adoption of standards and enforcing rigid naming practices and is based on migration of data. They are also looking at the length of time required for carrying out backups of records and how often these are being used to inform their plans for creating archival copies for the collections. File formats are more stable for text and images, but change more often for audio and moving images. In terms of storage media and conditions and updating metadata, the same strategy is followed for all media types. Quality control procedures in the life-cycle management include a checksum system (currently used by curators on the content), colleague review, archiving and use of standards. However, they are currently under review and are subject to change.

Their aim is to keep the digital deliverables available for as long as possible. The adoption of an exit strategy in the event of the institution being unable to meet the costs of sustaining the digital deliverables is an unthinkable question for the LC. On the contrary, the LC acts as the exit strategy for other institutions in the country. In their case, longer-term sustainability is not dependent on self-generating funds and moreover, the resource would not generate sufficient income to sustain it. The program has already secured partial resources for long-term sustainability.

12.6 Conclusion

General comments for others considering digitization projects, are the importance of clear goals as well as ensuring support from the highest level, taking into account the political aspect. In the case of the NDLP, the congressmen were very supportive as it provided a good argument to their constituents that their tax dollars were going towards something “good” and educational. This high level support also helped internally, to secure collaboration from all departments, even when they were not enthusiastic about the project. Since the targets set have been met and the popularity of the program and material created is obvious, even detractors cannot argue with the success of the project.

Few libraries actually have ongoing programs, rather than projects. Consequently, there is little comprehension of how complicated a procedure it is. Managers need to have some appreciation of the complexity of the digitization process. Due to the high costs and difficulties involved, extensive digitization projects are rare. There is now a great deal of information available on the web and elsewhere. There is no need to reinvent the wheel. Communication and collaboration are key in this field. With a project of this kind, knowledge of the material, the social history of the collections, library practices, and good technical and computer skills are all combined. It is necessary for staff to be flexible and willing to work in a team, and have good communication skills. Another point stressed was the importance of rapid generalization and standardization, trying to make all staff look beyond the specific great collection (especially for projects of this size and diversity). Consistency of metadata, for example, is very important. For smaller organizations, it is necessary to keep the project small and segmented, for example, choosing minimal cataloging or focusing on a small number of objects.

The LC and the NDLP are currently going through a period of transition. It seems that in the future, they will still carry out digital production but probably at a lower rate. They will carry out extensive work on the Digital Repository. The content has already shifted and will continue to do so - from solely Americana to a more

international content (links with Spanish and Russian culture are already being explored).

Generally, this is a case where the mandate was to produce a very large number of digital resources in a limited time at high quality, to be available to as many as possible. Although the target was very high, the fact that it was very specific from the beginning helped to move the team in one direction. The need for very clear goals shared by all in a project or program is a useful recommendation here. Quality control and use of standards in as many areas as possible were important principles of the program; a good example here is the way the Program handles and documents outsourcing and quality control of material returned from contractors. It is interesting to observe that the intensive production worked best with a flexible and informal management structure, with a group of committed and skilled staff working as a team, in a model that is different and less hierarchical from the rest of the organization.

valid xhtml 1.1
abp~04/02

13 Cornell University

HATII interviewed Peter Hirtle and Anne Kenney, from the Institute for Digital Collections and the Department of Preservation and Conservation at Cornell University, on October 18 2000. The Institute for Digital Collections endeavors to explore the application of emerging technologies, with a view to enabling greater access to cultural and scientific resources. As a result of their digitization projects, these resources are now available on the World Wide Web. In forming international partnerships with similar institutions, the Institute aims to amass a catalogue of digital collections from around the world. The Department of Preservation and Conservation was established in 1985, with the intention of providing resources for the research and teaching faculties within the academic community.

13.1 Organizational Digitization Program and Policy

There are three major areas within Cornell University Library currently engaged in digitization projects: the Cornell Institute for Digital Collections (CIDC), the Department of Preservation and Conservation, and the Albert R. Mann Library. The first two groups work closely together and are the focus of this report. CIDC is a component part of the Digital Library and Information Technologies (DLIT) division in the Cornell University Library. DLIT has responsibility for planning, implementing, and providing technical support for information technology programs and network linkages throughout the library system. The Department of Preservation and Conservation is also a component of the entire library system.

While both groups are involved in the digitization of both text and images, the primary focus of work in the Department of Preservation and Conservation has been on the digitization of textual materials (including illustrations in printed volumes), whereas CIDC has concentrated more on the digitization of visual objects. CIDC supports the middleware used for the delivery of materials it has digitized, relying on both homegrown database systems as well as software from Luna Imaging. Delivery of materials digitized by the Department of Preservation and Conservation is handled by DLIT.

13.1.1 Selection

Collection surveys have not played a significant role in the selection of material for digitization, however, the Department of Preservation and Conservation carried out a collection survey in the early 1990's. Instead, their approach to selecting collections for digitization has been to identify areas of known strength in the holdings. For example, one strong area at Cornell is its holdings of 19th century materials, which formed the basis for the Making of America (MOA) project. The Department will also scan on demand for faculty-related projects, such as distributed learning. The selection criteria include conservation concerns, teaching and learning potential, historical and cultural value, research significance, and the potential for increased access. The materials must have rich content to be selected.

Funding is the main factor in CIDC's prioritization and selection of material for digitization. All the projects within CIDC are funded on "soft" money. As a result, material from subject areas likely to attract funding has a greater chance of being selected. The CIDC supports projects initiated by Faculty members that will create resources for teaching. Other factors that influence selection are the teaching and learning potential of materials, the possibilities of enhancing access that may be created by digitisation, and research into digitization strategies.

The Mann Library has taken a radically different approach to selection. In both the Core Historical Literature of Agriculture project and the new Core Historical Literature of Home Economics project, the library has utilized panels of experts to identify the most important materials in each subject area. Those selections determine what is digitized.

Both CIDC and the Department of Preservation and Conservation have co-operated with libraries, museums, and archives, at all levels from local to international. While collaboration can be beneficial, both units feel that co-operation is not without problems and that firm rules for the collaboration have to be developed in order to produce a successful project.

13.1.2 Organization and Mission

Both programs are ongoing. CIDC was established in 1997 with lead gifts of $2 million from several benefactors. Other gifts have been received, as well as funding from the Provost and several Deans. CIDC, like the Department of Preservation and Conservation, also uses consulting income to support additional staff. While the initial gifts to create CIDC were generous, it was not enough to create an endowment. Regular funding would ensure the on-going maintenance of CIDC projects. Consultation work by both sections pays for more staff on a longer-term basis.

The Department of Preservation and Conservation has had digital projects in place since the early 1990s. The Associate Director of the Department also serves as Co-Director of CIDC, ensuring close cooperation between the two units.

The main purpose of both areas' projects is to create teaching and learning resources. In addition, both units hope to provide access to the materials to individuals and groups outside the library. The Department of Preservation and Conservation is also motivated by conservation concerns. Neither unit has revenue generation as a priority. Both have web pages that have explicit informative statement of intent, mission statements and further project information.

Between them, they have dealt with all types of material, with the exception of film, video recordings, radio broadcasts and TV broadcasts. The Department of Preservation and Conservation has had some dealings with sound recordings through the Music Library.

Both areas use standards, in particular TEI, JPEG, SGML and XML, MARC, EAD, AAT, and Library of Congress subject headings. Cornell feels strongly, that using standards will save money in the long-term. Where standards are not used, it is generally because they do not exist or are not appropriate to the project. Both programs started very early and many standards did not exist as such. This has led to Cornell establishing many of the accepted guidelines in the community.

The intended audience is similar for both programs. Both identify the Cornell academic community as its priority audience with other educational areas (such as K-12 education) meriting lower priority. The other main target audiences are museum, archives, and library users.

13.2 Project Management and Planning

CIDC does not carry out any pilot study as such, since it considers all of its projects to be experimental in some way. It applies the lessons learned from earlier projects to each new project, as well as scaling smaller projects to larger ones. The Department of Preservation and Conservation has conducted a number of research projects to answer technical issues relating to digitization, that must be answered before large-scale projects are implemented. It has also conducted usability studies on the user interface for MOA. The interview with University of Michigan gives more detail on this study.

Time and motion, planning, and scheduling tools are not used in either area.

Both programs digitize in-house as well as outsourcing digitization. In all cases, the nature of the material determines the digitization process; fragile or very valuable material is digitized in-house. Both programs believe that outsourcing can reduce costs. Specialized vendors, for example, can implement economies of scale. Pre-set cost figures for digitization can also insulate libraries from any increase in the cost of digitization, while the vendor has to absorb the increase. In-house digitization utilizes flatbed scanners (HP, Epson, Microtek, and Xerox), film scanners (Nikon Cool Scan), microfilm scanners, and high-end digital cameras (Phase One). Both programs use grayscale and color targets when scanning, and CIDC will often use tone balls when scanning three-dimensional objects.

Cornell University Library has established guidelines for digitization and data capture that are to followed by any project that hopes to deposit digital materials with the proposed central depository.

13.3 Human Resources and Training

CIDC employs seven regular staff, although additional staff can be hired to work on discrete projects. Smaller projects might be run by two to three staff, while a larger project may have as many as seven or eight. The staff

members tend to be humanists with an interest in technology. Projects often draw on staff from other areas of the library, such as catalogers and curators, to supplement the actual work of digitization.

The Department of Preservation and Conservation has 26 staff employed in areas such as administration and book conservation as well as digitization. These staff members tend to be librarians or conservators.

While some training is given to technical staff, the preference is to employ staff who already have good technical knowledge of digital processes. Staff are encouraged to attend the week-long Digital Imaging Institute course offered by the two units. They are also encouraged to take external courses when appropriate courses can be found. However, learning on the job remains an integral part of the training process.

13.4 Project Life-cycle Processes and Procedures

13.4.1 Intellectual Property

Both programs are aware of the copyright status of their materials. CIDC tends not to own the copyright on the materials it digitizes. When there is a question about the copyright status of a work, only on-campus or Cornell-affiliated users are permitted to print or download digital images, and only in direct support of teaching and research.

The Department of Preservation and Conservation generally only digitizes materials that are out of copyright in the U.S., such as those printed before 1923. It asserts a copyright claim on behalf of the Library, in the image headers. Neither unit uses watermarking to try to control further use of the digital files, relying instead on a general “terms and conditions of use” statement found on the web sites.

13.4.2 Preparation/Conservation

The Cornell University Library has a preservation policy that covers the handling of original materials from disbinding to scanning and reassembling. Material is not usually rejected for digitization because of its physical condition, but on occasion incomplete works may be rejected. As part of the digitization process, conservation staff will often replace missing pages, disbind, deacidify, wash, and mend the material to be digitized. Conservators provide training in handling to the scanning technicians. When disbinding cannot take place, the fragility of the original paper and binding presents the greatest challenge. Conservators assess the risk to original materials and help to identify the appropriate conversion methodology. Some scanning has been done using a face-up Minolta book scanner. Film intermediaries are also used when the risk of damaging the original outweighs the lower image quality derived from scanning an intermediate. In some cases, such as with the conversion of Art History slides, the only source material that exists is a film intermediate, which is then used for digitization.

Special equipment, such as cool lights, filters, and book cradles, is used to minimize risks. The processing staff and the curatorial staff work together to prepare the material before digitization.

13.4.3 Access

In general, once a digital surrogate exists, users are encouraged to use the surrogate. However, access to the original is not restricted.

13.4.4 Metadata

The starting point for the metadata in all digital projects is the catalog data. The nature, amount, and quality of the catalog data varies with the collection. Very often, the existing catalog data is enhanced with additional data created as part of the project. Links are created from the main catalog to the digital object through the use of an 856 link and a PURL server. For books, direct links from the catalog record to the book are created. In other cases, links lead to a finding aid encoded in EAD or to a collection-level record for a particular digitization project. Records for individual items are then stored in local databases developed in a variety of packages, including Filemaker Pro, Access, Informix, and SQL Server. Controlled vocabularies are used.

Information about the original object, the digital object, technical details, administrative information, and preservation-related information is captured at various points during and after the digitization process. The information is usually stored in either the image file header and/or local systems.

13.5 Formats and Compression of Digitized Materials

13.5.1 Formats

The Department of Preservation and Conservation has prepared several studies on the appropriate resolution and formats of digitized materials. TIFF is used in almost all cases as the file format for the master archival image. The format found on Photo CDs has also been used at the capture stage, but Cornell is in the process of making TIFF copies of many of its PhotoCD masters as a preservation measure. Many different kinds of formats are used when delivering images to the public, including GIF, JPEG, PDF, and MrSID file formats. Usually several sizes of images are available, from thumbnails through high quality versions. Finding aids are encoded in XML using the EAD DTD, and delivered to the user in HTML format.

13.5.2 Compression

Compression methods used include ITU Group 4, LZW, PhotoCD, JPEG and MrSID. Only lossless compression is used on master archival images, and not all master images are compressed. The aim of the compression of the master images is to decrease the amount of storage needed, while compression of access derivatives can improve the speed of delivery.

13.5.3 Quality Control

Some post-processing procedures are usually implemented. This can include color correction on calibrated monitors, the application of a sharpening mask, and contrast stretching. This is usually done using PhotoShop. CIDC and the Department of Preservation and Conservation both calibrate their scanners on a daily or weekly basis.

Quality inspection is carried out on 100% of objects. For visual images, this is usually done immediately after capture. For textual materials, inspection is usually done while adding metadata about the structure of each volume after capture. Checksums for the archival images are created so that any future error in transmission or file copying can be identified and rectified. There is no formal quality control for metadata. It tends to be done on an *ad hoc* basis, with errors corrected as they are found, but often the software used to enter metadata has functions to limit mistakes.

13.5.4 OCR

TextBridge is the software most often used to OCR text. Cornell feels that the level of accuracy achieved with TextBridge in the MOA project is comparable to that achieved by the University of Michigan. OCR is used to enable full text searching, not to create textual equivalents of the original. Hence, little of the OCR text has been corrected, and none has been rekeyed.

13.5.5 Access

Depending on the project, user access to the digital files can vary from restricted use by in-house staff, to worldwide access to the scanned images and metadata. However, the user is never allowed to have access to the original master TIFF image. There is no charge to use the resources.

Users can search on catalog and index data or, when available, on full text. They can often also browse by fields, and a thumbnail browse option is usually available. No special client is needed to access most collections; they are web accessible. Users of the visual collections accessible via the Insight software from Luna Imaging do gain extra functionality if they chose to access the images using Insight's Java Client rather than the web browser. The Java client enables users to create groups and slide shows, export images and data as web pages, measure features on images, and annotate them as well. It has a great level of interactivity and enables users to create very easily, their own resource from the larger body of work.

Information about the availability of digitized resources is disseminated through websites, press release, articles, print and broadcast media coverage, conferences and meetings, flyers and bookmarks. Locally, seminars for faculty are given to raise awareness. Web logs are used to monitor usage.

13.6 Evaluation, Funding and Long-term Sustainability

Both CIDC and the Department of Preservation and Conservation have conducted user evaluations of the different projects. CIDC investigated the use of its Utopia Project, a database of digitized versions of slides of the Italian Renaissance, while the Department of Preservation and Conservation in conjunction with Cornell's Human

Computer Interaction Laboratory, investigated use of the MOA collection. Based on its evaluation, CIDC selected Insight software from Luna Imaging to serve as the front-end for the dissemination of many of its projects. The Insight client addressed many of the issues raised in the Utopia project. The Department of Preservation and Conservation changed the selection criteria used in the MOA project after discussions with focus groups. *D-LIB Magazine* and *RLG DigiNews* have reports on evaluation strategies and user needs methods that are based in part on the Cornell evaluations. Front-end evaluation has become an iterative process and there are regular updates to project software.

Both sections would like to run evaluations that are more thorough in order to monitor user needs more closely. They would especially like to understand better, how users work with finding aids in particular and primary sources in general, in order to be able to develop better interfaces to those sources.

valid xhtml 1.1
abp~04/02

16 Center for Retrospective Digitization, Göttingen State and University Library

HATII interviewed Dr. Norbert Lossau, Head of the Center for Retrospective Digitization at the Goettingen State and University Library, on February 22 2001. In 1997, the German Research Foundation launched the funding program that led to the creation of the CRD. In its capacity as a digital research library, the Center contains collections of early printed materials, which are available on the WWW. The Center hopes to support ongoing activities within the GRF, while contributing to national efforts of standardization within the field of digitization.

16.1 Organizational Digitization Program and Policy

The Center for Retrospective Digitization in Göttingen (Göttinger DigitalisierungsZentrum GDZ) is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) and was established in May 1997 at the Lower Saxony State and University Library (SUB) in Göttingen, Germany. The GDZ is part of a funding initiative for retrospective digitization of Library materials building a German digital research library. As the Digitization Center of the Lower Saxony State and University Library Göttingen, it also receives financial support from the State Ministry for Science and Culture.

The GDZ collections are built around early printed book and journal collections such as North American travel literature (approx. 400,000 images), mathematical documents from the reference journal *Jahrbuch über die Fortschritte der Mathematik* (1868-1943) (250,000 images), and the Gutenberg Bible which, along with their accompanying structural and bibliographic metadata, are available over the world wide web. The main purpose of the GDZ is to acquire up to date know-how and experience in scanning early printed material, image processing, OCR and document management including access to digital libraries and books via the world wide web. The GDZ in Göttingen is also engaged in evaluation of tools and techniques for image capture and text conversion, bibliographic description, document management and the provision of remote access.

The GDZ is engaged in partnership projects like DIEPER (Digitized European Periodicals) where ten European countries have joined to build a virtual network and a central access point to make accessible periodicals that have been retrospectively digitized in Europe and around the world. It has recently started another similar project, Digizeit, which is intended to be similar in principle and aspects of functionality to the US-based JSTOR. The SUB Göttingen has received a further DFG grant to digitize in co-operation with 9 major research libraries 58 journals from 14 different academic fields.

Another important aim is to transfer knowledge into related library digitization projects and the coordination of national efforts towards standardization in various fields (e.g. digital conversion, online access, bibliographic description). Under the Retrospective Digitization of Library holdings funding program, the DFG also supports web-accessible resources held by other German research libraries and academic institutions (see http://gdz.sub.uni-goettingen.de/en/vdf-e/).

The priorities for digitization are set in the context of the overall initiatives in the historical research collections of the SUB, such as the building renovation projects to establish new reading rooms for 18th and 19th century physical materials with side-by-side access to the materials online. The collections policy centers on the Library's holdings of rare early printed materials, complete collections (e.g. the North American collections) and materials of wide appeal (e.g. the Gutenberg Bible). High priorities and selection criteria for digitization are research significance, enhanced access, use for teaching and learning, provision of user services, historical and cultural value, conservation and preservation.

Access to materials digitized as a result of projects aided by DFG grants is provided free of charge. Although income generation is not a high priority, the new Digizeit project, whilst remaining a not-for-profit project, will consider ways of generating sufficient income for it to continue past its first funded phase.

The primary audiences are firstly users of the University Library, mainly researchers, and a wide international

audience, particularly from the university sector in the USA. The GDZ regularly fulfills requests from US universities for documents from its collections which in addition to online delivery are delivered via FTP or CD. The other main audience is the general public worldwide, as some of the materials, such as the Gutenberg Bible, are of wide interest.

16.2 Project and Asset Management and Planning

The conversion process adopted by the GDZ follows the guidelines and recommendations of the technical working group implement funded by DFG outlined in its 1997 final report. Although there has been a technology shift since then, with scanning technology having moved on rapidly (for example, in 1997 it was not possible to use a bound volume scanner to scan in 600dpi), the basics of the strategy have remained sound. This strategy centered around the use of standards, both for metadata and for file formats.

Text material is scanned from microfilm (with new production date) and from the original. Scanning from microfilm (35mm) is contracted to a vendor; scanning from the original is undertaken in-house using the Library's own equipment. Images are created in 600 dpi.

The Zeutschel Omniscan 3000 scanner and the Minolta PS 7000 run under SRZ ProScan Book scanning software, developed for the GDZ by the Satz-Rechen-Zentrum Company in Berlin to meet special production scanning of older books (e.g. with TIFF-header editing, production control window with tree view over scanned pages, masking and cropping of pages during the scanning process).

Image capture of older books often requires some enhancement after scanning. For economic reasons this post-processing is done in batch mode wherever possible and black and white text material is compatable with this semiautomatic method. The GDZ uses a professional Norwegian program called PixEdit, originally designed for the CAD-sector.

In 1999 a new scanning device for capturing color images was added to the equipment of the GDZ. The digital Camera-Back Picture Gate 8000, max. resolution 8000 x 9700 pixels), manufactured by Anagramm GmbH, Germany, is used for face-up scanning of valuable Library resources (e.g. the Gutenberg Bible). Greyscale scanning of illustrations (e.g. from the travel account books) is also performed with this device. A special moveable cradle, first designed for Graz University Library by Manfred Mayer is used to ensure contact-free scanning for rare books. It utilizes a sensitive low pressure system to hold pages down.

The quality control criteria (e.g. legibility, completeness) and procedures for mass-produced text scanning are carried out with the software ACDC. For a complex and valuable project like the scanning of the Gutenberg Bible the procedures were much more rigorous because of the issues of color calibration which had to be closely monitored and thorough. Photoshop was used for color management.

Core bibliographic information is added to the TIFF header of the digital master to reference each image to a bibliographic record in the online library catalog. The digital masters are stored offline on CD-R, an ISO-standard storage medium.

16.3 Human Resources and Training

Human resource at the GDZ comprises a director, a technical director, systems administration and database manager, and specialists in OCR, digital imaging and a librarian. The GDZ functions both as a separate team carrying out funded projects (for example funded by the DFG, the European Commission or the National Science Foundation) and as a department of the Library which can carry out services for other departments. For example, there have been close collaborations with the Department of Rare Books and the GDZ on scanning of early printed materials and color digitization. The policy is to retain the GDZ personnel in the IT department of SUB if they are not working on GDZ projects, as the skills and knowledge of this highly specialist team are valuable to the organization.

16.3.1 The DMS – Asset Management

The overarching objective of the GDZ is to build up an open distributed digital research library, accessed via its new RDB-driven DMS, AGORA, which has become established as the central administration tool for the digital library of the GDZ. The software has been developed by a company in Berlin, Satz-Rechen-Zentrum (SRZ). It can be implemented on different RDB platforms and is based on an extensible document model developed at

GDZ. The DMS maps bibliographic and structural data (based on printed document structures) to a set of images, but it is open to handle any new document structures. The administrative functions of the AGORA system are controlled by an administrative tool which includes the following features:

- importing digital documents (scanned images and metadata in RDF/XML format);
- exporting digital documents in XML/RDF and PDF with structured bookmarks;
- batch converting high quality TIFF images to GIF, JPEG, PNG.

The largest part of the data model is used for documenting and retrieving digital resources and their parts. It is made up from a heterogeneous hierarchy of structural objects with common and distinctive features (e.g. monograph, title page, chapter), but special objects (e.g. figures, tables, indexes) can be referenced as well and it supplies document access via printed page numbers.

From the users' viewpoint the advantage of the DMS is that digitized documents are accessed exactly as a printed book might be accessed, e.g. via tables of contents, indexes, chapter headings, special features like figures or maps. The interface also preserves the printed page numbering (including searching in the printed page numbering schemes) and provides simple and advanced searching and a zooming facility.

The GDZ has chosen XML/RDF as the data interface of AGORA. The thinking behind this is that it provides good support for interoperability, efficient target-orientated retrieval by XML-compliant search engines, it is open for new document semantics, compatible to minimum Dublin Core and that it will become widespread and supplied by many software tools like editors and browsers.

In the future the GDZ will integrate full text retrieval (Verity) and user accounting support compatibility to other related metadata semantics and RDF schemas or XML DTDs.

16.4 Project Life-Cycle Processes and Procedures

16.4.1 Access to Digital Materials

Access is provided to both bibliographic records and digitized documents via two methods. Bibliographic records for the digitized documents are created in the PICA Online Library Network Catalog and serve as one method for providing access to the documents. Following the guidelines of the funding body, the DFG, digitized documents in Germany have to be recorded in the online library network catalogs (Verbundkataloge) to ensure direct access to a document entity via the online library catalog. The overall goal is the integration of the digital collections in an online library catalog (locally as well as in a networked catalog) to allow global subject searches over all library holdings independent of their physical representation (printed books, microforms, electronic publications, digitized books, etc.).

Göttingen treats digitized documents as reproductions (as microform – *Sekundärform* in German) and creates separate records in the online catalog. Special categories for online resources are added (e.g. Date of Digitization, Creator, Copyright, URL). The availability of the PICA-GBV Online Library network catalog on the world wide web allows users to start a search in the Online Catalog and go directly from the hitlist to the electronic version. The bibliographic description in the GBV catalog conforms with the German exchange format for libraries (MAB-2).

Another method of access is directly via the Document Server located at the Library's homepage. Simple and advanced searching as well as title/author browsing for collections provides easy access to digitized documents. Users have the option to view the document on-screen via a standard web browser (as GIF, JPEG, PNG images), download a PDF file for printing, or order a high quality (600dpi) printout from the Library. Additionally, a PDF version, together with the free Acrobat Reader, will be offered on CD-R. These versions for output will be generated as derivatives from the digital master.

16.4.2 Rights

The copyright on materials is owned by the Library and material is made freely available with few exceptions on use apart from use in broadcasting.

16.4.3 Handling

The GDZ, in its collaborations with the Department of Rare Books in the SUB has built up a significant body of experience in the handling of early printed materials during the digitization process. For further details see: Norbert Lossau & Martin Liebetruth, "Preservation Issues in Digital Imaging Technology" in *Microform & imaging*

review, K.G. Saur, Vol. 29. No. 4, 2000 (originally published in *Spectra*, Museum Computer Network, Fall 2000, Volume 25, Issue 2, S. 30-36).

16.5 Format, Resolution and Compression of Digitized Materials

16.5.1 Text

Text-based materials are scanned as TIFF images. A well-known problem in OCR is the recognition of textbooks in *fraktur* (gothic). As the broad evaluation of OCR-programs by the GDZ pointed out, standard programs as well as sophisticated trainable programs (Prime Recognition, ProLector, Optopus, FineReader) do not provide solutions for an economically automatic recognition of these kind of texts. This is a serious problem for many digitization projects in Germany. The GDZ in Göttingen uses the Russian program FineReader (vers. 4.0). Non-gothic texts, even from older books and difficult image quality can be recognized with a low failure rate, allowing the creation of a text index for background-search.

The GDZ is looking for a technical solution for the economic recognition of gothic texts by co-operating with a company derived from an academic institute of Mathematics (in Potsdam,Germany). Other institutions in Germany also discovered this problem as a chance to fill a special segment of the market for text recognition programs. For an intervening period – until the issue of Gothic text processing can be resolved – the minimum for accessing digitized text will be the availability of navigational tools given by the original itself, like tables of contents, indexes, list of illustrations, hyperlinked to the referenced portions (e.g. chapters, image pages) of the documents. Materials using gothic text are sometimes sent to a service vendor for re-keying, but this is mainly only done for tables of content and rarely for full texts.

A new project funded by the European Commission and led by the University of Innsbruck in Austria is researching the development of a modular addition to the OCR program Finereader to recognize gothic characters. The GDZ has shared experience with the research team and is working in partnership with them.

16.5.2 Images

Images are created in 600dpi. The scanning process produces a high quality digital master in TIFF-format. Derivatives (GIF, JPEG, PNG) are created on the fly for online delivery. A PDF version can be obtained for downloading or for offline delivery on CD-R, burned by the GDZ.

16.6 Evaluation, Funding and Long-term Sustainability

No formal evaluation has been carried out with users, but the unsolicited user feedback has been good. There has been an especially good level of response from users in the USA, with many commenting positively on the free availability of materials.

The longer term funding position is to continue to pursue funded projects for the GDZ such as Digizeit, and to retain personnel. The GDZ is now at the end of its second two-year funding tranche from the DFG. The GDZ will continue even if further external funding is not forthcoming and, as noted above, absorb staff into the IT department of the SUB in order to retain the skills and experience that have been built up.

However, the outlook for funding is promising. The technical infrastructure and the knowledge base (particularly in color digitization and handling early printed books) at the GDZ is so strong that it is felt that there will always be an opportunity to generate income from digitization consultancy and services to clients such as other libraries and archives with valuable or rare/early printed materials to digitize.

The strategy for longer term sustainability, data preservation and archiving is to follow standards in all fields and to keep preservation copies of all data. For example, master TIFFs are always kept for graphic formats. These can amount to considerable amounts of data: for the Gutenberg Bible there are more than 700 CDs of data as one illuminated page might take up 400 MG of storage as an uncompressed TIFF file. CD-R, an ISO standard storage medium is used. Similarly, metadata is stored as ASCII structured data.

The thinking is that there will have to be migration of data in the future to other formats, but the use of standards now will facilitate that migration. The GDZ strategy started out with adherence to standards and it has proved to be a sound foundation for the Center.

valid xhtml 1.1
abp~04/02

17 Harvard University, Visual Information Access (VIA)

On January 26 2001, HATII interviewed Martha Mahard, Curator of Visual Collections at the Fine Arts Library of Harvard University, and Ann Whiteside, Visual Resources Librarian at the Frances Loeb Library of Harvard's Graduate School of Design. The Visual Information Access project is co-ordinated between Harvard's Fine Arts Library and the Graduate School of Design. The aim of the project is to create a comprehensive catalogue containing information about the materials and artifacts housed in the University's libraries, museums and archives. The materials are digitized for the purposes of preservation, increased accessibility, as well as to serve as resources for teaching and learning, for users both within and outside the University community.

17.1 Organizational Digitization Program and Policy

The Visual Information Access (VIA) system is a union catalog of visual resources at Harvard University. It includes information about slides, photographs, objects and artifacts in the University's libraries, museums, and archives. This system represents the first phase of an ongoing effort and is frequently updated. The program started in fall 1996, completed phase one in April 1998, and is ongoing. The participating libraries, museums, and archives are:

- Arthur and Elizabeth Schlesinger Library on the History of Women in America, Radcliffe Institute
- Baker Library, Harvard Business School
- Harvard Fine Arts Library, Visual Collections
- Harvard University Art Museums (Busch-Reisinger Museum, Fogg Art Museum, Sackler Museum)
- Harvard-Yenching Library
- Loeb Library, Harvard Design School
- Peabody Museum of Archaeology and Ethnology

17.1.1 Collection Survey and Priorities for Digitizing

The institutions that participated in VIA conducted a collection survey which examined images holdings, how much information was available in automated databases, etc. This was useful to confirm their initial ideas; it gave a clearer indication of the breadth and scope of the initiative and showed which institutions wanted to be automated. The administration had very limited understanding of the size of resources and the implications of the project. The survey could have been more detailed, but it focused more on data collection and management issues, which was still very useful.

The collection survey was also used to establish priorities for digitizing holdings. Initially those with automated collections moved up the prioritization list. After that, the Library Digital Initiative (LDI) forced prioritization with the grants they offered.

The individual repositories make their own prioritization decisions. Many people digitize the material that is likely to attract grants, rather than what they would have preferred. External forces, such as the LC/Ameritech grant, allowed them to digitize Americana at the Graduate School of Design. This would otherwise not necessarily have been high on the prioritization list. Selection is also influenced by what is taught at the curriculum.

The Task Group Report on Visual Resources at Harvard, produced in May 1997 acts as a kind of strategic document.

The objective of the digitization policy was to create a union catalog to help facilitate access to visual resources (whether there were images accompanying the catalog or not). The catalog was to act as resource discovery tool. The project has been successful in achieving its aims. The time was right for VIA, as Harvard was exploring ways of offering access to visual resources beyond books. People now want to put their material on VIA.

One of the lessons learned from this process is that it is useful to make people aware that curatorial ideas and wishes do not always match what the funding will support.

17.1.2 Overall Obstacles

Copyright is a serious impediment. One of the greatest obstacles was to establish use of a common language. Many participants did not have the technology at their fingertips. There was a variety of different needs, from the Museum, which wanted the highest possible resolution to the Schlesinger, which wanted the lowest. There was an overload of technical information that they had to learn. Communication with technical staff was not always easy. Another difficulty was setting standards and realizing that they could not have a single benchmark across the board. There were no sufficient or appropriate guidelines available when they started, so they did not know what constituted an acceptable deliverable.

17.1.3 Selection and Prioritization of Materials

The main criteria for selection, used by most of the individual repositories, were primarily intellectual property rights and enhanced access, followed by teaching and learning potential and conservation. These changed to some extent over time. For example, preservation was also an initial consideration, but they are now more concerned with the limitations of digital information.

17.1.4 Co-operation

There was co-operation with archives, other libraries, museums, and academic institutions, whese were all local.

This experience showed that in any collaborative project, it is important to be receptive to others, in order to learn from each other. It is important to try to get managers and technical staff on board first and to allow time for collaboration, since building communication takes time. Email helped, especially to discuss sensitive issues and disagreements that could not be mentioned in face-to-face meetings.

17.1.5 Purpose of the Program

The digital deliverables are created for preservation, public access, teaching and learning resource, research, and wider access. The program produced a statement that was explicit about its scope and rationale. This is available on the web (http://via.harvard.edu/html/VIAscope.html) and refers more to the type and size of material contributed by each partner. This complements the brief statement at the opening VIA web page (“The Visual Information Access (VIA) system is a union catalog of visual resources at Harvard … This system represents the fist phase of an ongoing effort and additional information will be added on a regular basis.” http://via.harvard.edu/html/VIA.html).

17.1.6 Nature of the Source Materials and the Impact of Digital Deliverables

The program digitized a wide range of source material with varied format and nature:

- printed illustrations;
- artifacts;
- original artworks;
- intermediaries (film, microfilm, microfiche, photocopies etc.);
- photographic negatives;
- photographic prints;
- textiles.

The materials selected for digitization were both the entire body of some collections, as well as representative samples. The digital deliverables were not intended to be re-used or re-purposed, but for some projects it was the other way round. Projects like MESL for the Museum and LC/Ameritech for the Design Library allowed them to digitize material that they brought to the program. This did not seem to work very well, with a great deal of massaging of data, rescanning and reformatting of older images. Another problem was the amount of material on Photo-CD that needed to be copied and that also demanded resources. If they had unlimited resources, they would have copied high-resolution images on the server. Problems such as these made them realize that it is usually not possible to find the perfect solution, for example to scan perfectly once. It is therefore useful, to accept that you are scanning for today and not necessarily for the future and that you might have to rescan in the future.

17.1.7 Interoperability

For representing content they used JPEG and SGML. For describing content, they used MARC, Dublin Core, Categories for the description of Works of Art, and the Visual Resources Association Core Categories. For controlling data values they use AAT, Library of Congress subject headings, Library of Congress subject thesaurus, and name authority files, but these are not imposed.

They looked at existing guidelines for digitizing particular document and object types, but did not find any that were relevant or useful.

They realized the necessity of compromise and accepting from the beginning that there is no one right answer. That is acceptable as long as everyone agrees on one standard, so that everyone may claim joint ownership of the project.

17.1.8 Target Audience

The primary intended audience for the digital deliverables is: four-year college, graduate school, distance learning, public library and archive users. The vast majority of resources on VIA are accessible to all, so apart from the targeted audience, anyone who desires to use Harvard's public catalogs and the linked objects that have not been restricted to the Harvard community, can do so. This may be of interest to school teachers.

The program did not acknowledge the needs of those with disabilities.

They are working on security of the system and access restrictions at the moment. Restrictions are due to copyright and sensitivity of material.

Evaluation of the targeted audience is going to be carried out as the next step of the program. Five or six people on the central team have been working in this area for many years. They need independent staff to carry out the evaluation, since those working on the program are too closely involved and lack the right specialization. Although they do not have any detailed hard data about the system's use, there are more users than were anticipated.

17.2 Project Management and Planning

They received project management advice from the wider University. The management of the program is carried out by the Steering Group with help from the staff of the Information Systems department of the central Library. The collaborative nature of the program has led them to explore other ways they can share funding. They have regular meetings of the Steering Group, an email list, and regular checks by the Systems Office. Initially they did things individually, without delegating, which did not work. These project management procedures were suggested to them by the University Library Council to which they report. They did not carry out any feasibility or pilot study to assist with the planning of the program.

Digitization was mainly carried out in-house because the central Library Digitization Initiative (LDI) wanted to develop local infrastructure and skills. In some cases it was outsourced, because of historical legacy with the CD-ROMs.

Technologies Used for Image Digitization
Flatbed scanners (various manufacturers)
Film scanners
High-end professional cameras

They established a set of guidelines for document scanning and digital photography (see LDI report). They used grayscales, color charts, and targets as benchmarks (see LDI and Museum report).

17.3 Human Resources and Training

The number of staff who work on the VIA program are as follows:

Type of Staff	Number	% of time on the project
Director (Steering Committee		
Metadata specialist (1 per team)	1	30
Curator	10	30
Digitizer	2	50
Photographer	2	50
Technical support staff	2	50 +30
Technical development staff	3	30

The background and profile of most people on the project is curatorial and technical. Only one member of staff was re-deployed from another area.

Although advice was available in-house about technical aspects of digitization, they also used some external consultants (e.g. Eastman House Permanence Institute). Areas where training needs have been identified include application of technical standards, preparation and handling of materials for digitization, technical operation of digitizing equipment, post digitization processes, metadata creation, and digital preservation. Training was received by curatorial staff, specialist technical staff, library/cataloging staff, and equipment operators. This was organized in a variety of ways: in-house, using project staff, the Library's own consultants, and external consultants, by attending external courses, with independent study, and by learning on the job. The training successfully met the needs of the organization.

17.4 Project Life-Cycle Processes and Procedures

17.4.1 3.1 Reproduction and Copyright

They are aware of the copyright status of the digital deliverables that the program is creating. Due to the nature of the material (e.g. with artists often owning copyright of analog works), this is a complex area. Although they own the copyright for most of the materials that they digitize, there are some where they do not, but these can be used via the standard legal provisions made for libraries. The rights information is available, but they do not put it on VIA. There is a standard copyright statement on the VIA opening web page: "Most of these materials are

owned, held, or licensed by the President and Fellows of Harvard College. They are being provided solely for the purpose of teaching or individual research. Any other use, including commercial reuse, mounting on other systems, or other forms of redistribution requires permission of the appropriate office of Harvard University". Users of VIA can make printouts on paper or download to a PC both thumbnails and lower quality images. No electronic management systems are used to control copying.

17.4.2 Preservation/Conservation

There is a conservation procedure for original materials. The investigations carried out into the condition of the original materials prior to digitization vary according to the institution participating in the project. No risks to the material were identified during the preparation or digitization process. For some material they specified special equipment and digitization processes, e.g. special handling, use of cradles. The material is prepared/monitored by curatorial and conservation staff before and during digitization. In most cases, the organization will not restrict access to the original once it has been digitized, except for preservation reasons, e.g. lantern slides.

17.4.3 Preparation and Sources

The cataloging and reference systems used by the VIA participating institutions before digitization were a variety of both manual and automated. For example, the Museum uses EmbARK, while the Library has created a customized version of EmbARK. Others use FileMaker Pro. In some cases material had to be altered from the original form for the digitization process. No material was rejected before digital imaging. They digitized originals, as well as reproductions and intermediaries (the majority). They used 35mm slides, 4x5 or 8x10 transparencies, and photographic prints.

17.4.4 Metadata

The originals are in most cases cataloged. The metadata standards used to catalog originals are based on the Harvard core, their own in-house system which is developed using the SiteSearch database system from OCLC (the OnLine Computer Library Center) and uses Access as a front-end. Some of the participating institutions also use Dublin Core, MARC, VRA, CDWA. They use a controlled Vocabulary, thesaurus and a classification scheme to control data values. The metadata records information about the original object, the digital object, the digitization process, technical details, staffing details, and administrative information. The metadata records were created by a range of staff, including digitizers and archivists/information professionals. The metadata records for digital deliverables are stored in a separate catalog, which is available on the web.

17.5 Format, Resolution and Compression of Digitized Materials

17.5.1 Images

The file formats used for images were JPEG and TIFF for delivering, and TIFF, Photo-CD and JPEG for capturing. The resolution used is 300-600dpi for capturing and preserving and 800 pixels (dpi varies) for delivering images. They used 1 to 48 bit depth for capturing and preserving, and 4 to 24 for delivering. Group 4 PCD compression was used at capturing and delivering stage while JPEG compression, GIF/LZW was used for delivering. The aim of the compression was to reduce cost (by reducing file size, since storage in Harvard's Digital Repository is charged annually by GB), to improve access, to enhance usability, and to decrease storage requirements. The original scans are retained in uncompressed form, except for 1-bit scanning and Photo-CD scanning projects). See LDI report for details.

17.5.2 Quality Control

The quality of digital deliverables rests with each repository, but they are currently discussing issues around that at this stage. They carry out a random check on materials. The quality control for metadata recording relies on checking with the catalogers.

17.5.3 Delivery, Access and Use of Digital Deliverables

Users have free open access to both the catalog and the digital deliverables for browsing, but not for manipulating them in any way. This is high on their list of priorities for exploring in the future. Access to some material is restricted to in-house users. Users can browse by 9-10 fields (e.g. name, title, subject, place, nationality, date, repository, ID). They monitor use via automatic data capture by the digital repository and other systems and web logs. The level of use has not been very high yet (under 2,000 hits per month), but is growing.

17.5.4 Dissemination and Publicity

The methods used for informing potential users about the digital deliverables created are: announcement on website, press release, articles in print media, announcements at conferences and meetings. This is a very important area, which they will develop, as VIA is still underused.

17.6 Evaluation, Funding and Long-term sustainability

They have not carried out any formal evaluation yet, but are about to organize one now, as mentioned above in the section on target audience.

Each repository contributed the cataloging and the staff. The Office for Information Systems of the Library contributed the rest. They produced an interim preliminary report.

All elements of the project will need to be updated. New materials and metadata will need to be digitized all the time, metadata will need changing, the user interface will probably need to be changed, and it is highly likely that they will have to change file formats. They are in the process of developing a preservation strategy. No exit strategy has been established.

17.7 Conclusion

General comments about the program included the need for curators to be able to communicate better with technical staff or try to find “interpreters” initially. People starting in this area should not think that they can do it alone. Administrative support and funds, but also belief in the project are all key. It is important to secure institutional commitment. They were fortunate that they did this at institutional level.

In this interview, it was interesting to see the perspective of two committed librarians / subject specialists who embarked on a digitization project that has since become a program. They had to learn a lot of technical information and a new language, and overcome a steep learning curve. The digitization activity was an add-on project for their institution and their own time. In this type of collaborative project where several institutions from a wide range of disciplines are working together, it was evident that communication was vital, as was stressed several times during the interview. Setting common standards across the field was also important, while accepting that no one solution will work for all materials. The wide-ranging nature of the content and format of the material that is being digitized is an interesting model for the provision of access to visual resources that other universities and similar institutions could benefit from, with serious implications for teaching and research.

valid xhtml 1.1
abp~04/02

18 Harvard University Fine Art Museums / Collection Management System

On January 25 2001, HATII interviewed Andrea Notman, Registrar for Collections at Harvard University Fine Art Museums, and Peter Siegel, Director for Digital Imaging at Harvard Fine Arts Museums and the Fine Arts Library. The Collections Management System was established to support initiatives within the Museums, such as exhibition programming, internal research interests and educational schemes. The digitization of the source materials not only allows for increased accessibility and cross-fertilization between departments in the Museums, but also enables more fragile objects to be accessed without risk, as well as being an effective method of cataloging new acquisitions.

18.1 Organizational Digitization Program and Policy

The Harvard University Fine Art Museums include the Fogg Art Museum (Western art from Middle Ages to present), the Busch-Reisinger Museum (art of German-speaking countries), and the Arthur M. Sackler Museum (ancient, Asian, Islamic, and Later Indian art). Their first efforts with digital imaging started in 1993-4 and they are currently moving into an electronic collections management system which supports digital images. This is an ongoing program.

18.1.1 Collection Survey and Priorities for Digitizing

At the time of the interview the Museums did not have a formal digitization strategy, but were considering developing one. They carried out several discussions about digitizing the different types of collections, rather than a formal collection survey.

Factors that influenced priorities for digitizing were risk management for objects in the collections, exhibition-related activities which guided selection, and cataloging of new acquisitions. They see the digital collections management system as a tool to support day-to-day activities in the Museums, a resource for staff, and a way to overcome physical restrictions of use of the collections. Selection is generally internally-driven by exhibition program, staff research interests, and education activities.

Their experience highlighted the importance of having the collections management system established prior to embarking on digitization. They also observed, that all the layers of documentation are important and should be recorded and preserved. Their experience showed that high-level managerial support is vital.

18.1.2 Overall Obstacles

Some of the obstacles to planning the development of digital deliverables were: assessing the collections and talking to curators; overcoming institutional resistance to change and a high learning curve (e.g. with MESL).

Obstacles to building digital deliverables were: infrastructure problems; the late creation of Harvard Library central Digital Repository; and large data storage needs.

18.1.3 Selection and Prioritization of Materials

Criteria used in the selection and prioritization of materials for digitization include:

- teaching and learning potential;
- conservation (decreased wear and tear);
- historical and cultural value;
- research significance (ability to assist research projects);
- provision of user services;
- enhanced access;
- potential to reach disadvantaged groups (more a future issue);
- research into digitization strategies.

A further criterion, although not in our questionnaire list, was to support collections management.

Intellectual property rights are not a selection criterion, since they might digitize material they do not own the rights to, for purposes of internal documentation.

18.1.4 Co-operation

Although they collaborated with various institutions and organizations, all these were in the periphery of their digitization activities.

Generally they would recommend that you make sure you "clean house" first, before you start collaborations, so you can start from a position of strength.

18.1.5 Purpose of the Program

The purpose of the Digitization program is to build digital collections primarily as a teaching and learning resource, to support research, and assist preservation by reducing wear and tear. Another purpose is to assist in the production of catalogs and publications. They have not produced a formal statement of intent for their digitization activities, which support their basic aim to document the collections.

18.1.6 Nature of the Source Materials and the Impact of Digital Deliverables

The program digitized a range of visual material with varied format and nature:

- printed illustrations;
- handwritten documents;
- artifacts;
- original artworks;
- intermediaries (film, microfilm, microfiche, photocopies etc.) (up to 8"x10");
- photographic negatives;
- photographic prints;
- textiles.

The materials selected for digitization were intended to comprise the entire body of some collections. They considered the question of misrepresenting the collection when only a part is shared with the public. As a general rule the digital deliverables are intended to be re-used, as their main principle is "scan once, use many".

18.1.7 Interoperability

For representing content they used the Dublin Core DTD, JPEG and XML. They also use the NISO (National Information Standards Organization) Image Preservation metadata (http://www.niso.org/ commitau.html). For describing content, they used the Categories for the Description of Works of Art. For controlling data values they use AAT, Library of Congress subject headings for some photographs, and their own system. There is a lack of documentation about these types of collections, e.g. geographical place names. When organizations develop their own system, they must standardize and consistently document it.

They looked at existing guidelines for digitizing particular document and object types, but did not adopt any. Few were available at the beginning. Although this situation has changed, a higher level of detail is still needed when adopting guidelines within the institution.

They realized that one specification cannot work for all users. The technical specifications vary according to source matter. These are all different from pre-LDI specifications.

18.1.8 Target Audience

The primary intended audience for the digital deliverables is Harvard University students (four-year college and graduate school), as well as Museum users. Audiences wider than the one intended can also share the parts that are publicly available.

18.2 Project Management and Planning

They used both in-house and external advice on managing digitization. Digitization activity is a sub-component of the larger organization of the Fine Art Museums. It led to changes and re-evaluation of organizational relationships and procedures. Initially they had the VIA Steering Committee as a formal project management

structure, before Peter Siegel was hired to head digitization initiatives.

The QA procedures they use are: ISO image creation and viewing; checking integrity of data; curators vet the materials created; press proofs comply with CMYK at the press houses; use of image metadata; and use of the MD5 checksum algorithm for ensuring the quality of digital files.

They carried out a pilot study digitizing 300 images that assisted them in scheduling, definition of training needs, technical feasibility, user needs, workflow analysis, workflow piloting, and technology forecasting. A small change from the pilot was that conservation now plays a very active role when they start a project. They used Gantt charts, but found that they are ineffective when running five projects simultaneously. The best option is to choose professionals suited for the job in the first place.

Digitization was carried out in-house because of the quality control that they required and because the collection needs to stay in-house. They purchased all the necessary equipment.

Technologies Used for Image Digitization
Flatbed scanners (Heidelberg, Topax iX)
Drum scanner (will buy one for collection of X-rays where it can penetrate the density) & for old negatives
Film scanners
Digital cameras
High-end professional cameras (Phase One)

They established a set of guidelines for data capture procedure and specified color management guidelines. They use grayscales, color charts, SR37 target for black and white 2D calibration, and digital photometers for monitors. They scan raw color proof. One problem is that most commercial targets are designed for synthetic material.

18.3 Human Resources and Training

The number of staff who work on the program are as follows:

Type of Staff	Number	% of time on the project
Director	1	100

Programmer (extracts metadata automatically)	1	100
Curator	9-10	5-10
Digitizer	1.5	100
Photographer	3.5	100
Technical Support staff	5	10

Most people have a museum background or are industry professionals.

Areas where training needs have been identified are: project management; application of technical standards; preparation and handling of materials for digitization; technical operation of digitizing equipment; post digitization processes (especially color management and preparation for press); and metadata creation. The photographers received appropriate training.

Training has been organized in a variety of ways, in-house using project staff, by attending external courses, with independent study, and by learning on the job. The training did meet the needs of the organization. In addition to this, they attended more industrial-type training, offered by the camera manufacturers.

18.4 Project Life-Cycle Processes and Procedures

18.4.1 Reproduction and Copyright

They are aware of the copyright status of the digital deliverables that the program is creating. Due to the nature of the material (e.g. with artists often owning copyright of analog works), this is a complex area. Although they own the copyright for most of the materials that they digitize, there are some where they do not own the copyright but have a license to use the material. The rights information is available and they will also declare the copyright status of the digitized material in the future. Users of the material can make printouts on paper or download to a PC both thumbnails and lower quality images as long as it is for fair use. Harvard users can also download the high quality images. No electronic management systems are used to control copying.

18.4.2 Preservation/Conservation

There is a conservation procedure for original materials. The investigations carried out into the condition of the original materials prior to digitization vary according to material and include conservation survey and stabilization. Risks identified during the preparation or digitization process include heat, especially from lighting. They put UV gels in front of lighting sources and use special equipment. The material is prepared by curatorial and conservation staff before and during digitization. They will only restrict access to originals once the material has been digitized for conservation reasons.

18.4.3 Preparation and Sources

The cataloging and reference systems used are EmbARK and 3D System. The information used from these systems in the digitization process included dimensions, material, title, and artist. They digitized from originals as much as possible, but also used reproductions and intermediaries. They used 35mm slides, photographic prints, 4x5 and 8x10 transparencies.

18.4.4 Metadata

The originals are cataloged using EmbARK. Eventually the same system will be used for cataloging the digital surrogates when the Library's Digital Repository Service is fully operational. In the meantime, they record Dublin Core metadata for the digital surrogates using a spreadsheet. They record metadata about the original object, the digital object, the digitization process, technical details, staffing information, and administrative information. The metadata records are created by archivists. The catalog for the digital deliverables is available on the internet.

18.5 Format, Resolution and Compression of Digitized Materials

18.5.1 Images

The file formats used for images are JPEG for delivering, and TIFF for capturing, preserving, and delivering. The resolution used is 6Kx7K for capturing, preserving, and delivering images. They used 48 bit-depth for capturing and preserving, and 24 for delivering. Compression is used at delivery stage. The aim of the compression was to improve access and enhance usability. They retain the original scans in uncompressed form and without color matching. Post-scanning processing is carried out on the images, mainly color correction, sharpening of low-resolution images which is done automatically with a script but is visually inspected. The average file sizes created are around 200 MB at capturing and preserving stage, and 2Kx3K for delivering. They use specialized targets and software to measure performance of digital cameras and scanners (e.g. Kodak Q-60 color input target). These same targets are used to generate ICC (International Color Consortium) profiles for color management. They check the dynamic range of their equipment with an MTF target and Delta E (which measures changes in color).

Their experience showed that you should try to make a custom profile for the equipment, document it, and stick to it. For publications, you also need to talk to the publisher about the needs for the publication, talk to their technicians and the academic community to determine what the collection requires. Color management is imperative after capture.

18.5.2 Quality Control

They carry out a total check to ensure the quality of the digital deliverables. Quality control procedures might require a little more time at the beginning, but save a lot of time in the long term.

18.5.3 Delivery, Access and Use of Digital Deliverables

Users have free open access to both the catalog and the digital deliverables for browsing, but not for manipulating them in any way. More advanced ways of using the material will evolve. Access to some material is restricted to in-house or Harvard University users. They monitor use via automatic data capture.

18.5.4 Dissemination and Publicity

At the moment no publicity is used to publicize the creation of digital deliverables. In the future, the methods that they will use might include: announcement on website, press release, articles in print media, announcements at conferences and meetings, and announcement at electronic listservs.

18.6 Evaluation, Funding and Long-term Sustainability

They carried out front-end and formative evaluation using focus groups and hands-on testing. They also evaluated informally internally, with other museums and with the publishing industry, which did not lead to major changes, but helped them to confirm a lot of their ideas and raised their comfort level. For example, suggestions from the press industry led them to keep the color bars in the archival image.

The program was funded at $1 million from internal Harvard University funds for the largest part (three-quarters) and about one quarter from private funds, e.g. from the Getty, and the Carpenter Foundation. They are provided sufficiently, but could use more space. They think that using standards saved them money in the long term. There are regular updates for the funding bodies about the progress of the program. They also had to provide a collection survey, a conservation plan, risk assessment, front-end evaluation report, project management plan, cost models, and workflow reports.

All elements of the project will need to be updated. New materials and metadata will need to be digitized fairly regularly; metadata will need regular updating; the user interface will need to be changed; and it will be necessary to change file formats probably within the next two years. They have a preservation strategy that includes file formats, storage media and conditions, and metadata updates. The strategy is based on migration of data. Quality control procedures in the life-cycle management are based on the MD5 checksum algorithm which checks the quality of digital files. Longer-term sustainability is not dependent on self-generating funds. They have not adopted an exit strategy.

18.7 Conclusion

As general comments, they mentioned that it is very important to keep appraised of professional industry developments, not just academic ones. Digitization programs require a very professional approach, and must be taken seriously, as they are not for the faint-hearted. It is necessary to talk to colleagues, and establish support groups.

An interesting aspect of this case is that due to the nature of the materials, which is a wide range of artworks, and the nature of activities that digitization might support, such as professional publications, great importance is placed on color management, calibration of equipment, and the adoption of procedures and practices across all steps of the digitization process to ensure faithful color reproduction and high quality images. For this reason, the type of equipment used is at the high end of the market and there are also very large file sizes. In this case the range of uses that digital images will be put to is very wide, going beyond simply viewing on the web towards use as a tool to support all aspects of the Museums operation. Another characteristic was the emphasis on the importance of links with professional publishing and the industrial world. They also mentioned that in several cases they had to design their own custom profiles for the equipment or develop their own system for controlling data values. Their view is that this is acceptable as long as it is standardized, applied consistently and is documented.

valid xhtml 1.1
abp~04/02

19 Harvard University Library / Library Digital Initiative Program (LDI)

HATII interviewed Wendy Marcus Gogel, the Digital Library Project Liaison at Harvard University Library, and Stephen Chapman, the Preservation Librarian for Digital Initiatives, on January 26 2001. The Office for Information Systems at Harvard University, conceived of the Library Digital Initiative as a means of creating a comprehensive infrastructure to manage the digitization activities of the University. In the creation of a union catalogue, the Initiative seeks to address issues of acquisition, organization, delivery and archiving of digital deliverables within the library, with the resultant material being aimed principally at the University community.

19.1 Organizational Digitization Program and Policy

Harvard University launched the Library Digital Initiative (LDI) in July 1998 as a five-year program to develop the University's capacity to manage digital information by building infrastructure and expertise in this area. The three main areas of program activity are: Advisory Services, Technical Development, and the Grant Program. One part of the program is based on an internal bidding process, the Internal Challenge Grant Program, open to all parts of the University. This decentralized model for building content follows Harvard's tradition on building library collections, keeping decisions about what to collect at a level close to researchers and students. The program is managed under the direction of a Steering committee of representatives from faculties and libraries. This is responsible for setting priorities for program activities, for monitoring progress and costs, and for allocating the incentive fund. The selection process for the Internal Challenge Grant Program is managed by the Grant Review Committee. Each project that has been awarded a grant has its own project manager who collaborates with the central Library staff. Because of the different projects that are part of the LDI, there is variety in the specific digitization approaches followed, depending on type of material, staff involved, purpose, and particular circumstances. Therefore, although they act as both developers and advisers at the LDI, they are not a centralized authority. They have learned what people want to control and what they would like to pass on centrally (e.g. digital preservation) and are focusing on these areas, operating on a consensus basis.

19.1.1 Collection Survey and Priorities for Digitizing

The individual projects that are bidding for an internal LDI grant are encouraged to carry out a collection survey. The way this is carried out and the level of detail recorded vary between projects. In practice many projects based their methodology on a very small sample of the material. As a result, some work plans were based on insufficient information and additional analysis and work was required.

The LDI Grant Review Committee is responsible for setting general priorities through the Grant Program, while the LDI Steering Committee may set strategic initiatives, such as the access projects specifically intended to provide enhanced access to collections by adding information to Harvard's various catalogs. The requirements explained in the Call for Proposals for the Internal Challenge Fund specify that successful proposals require institutional support, collaboration with LDI staff, library sponsorship for proposals submitted by non-library units in order to ensure access to information, and project management from the project unit. The selection criteria are based on the nature of the materials, the expected level of use, the level of sponsor commitment and contribution, the contribution to the LDI collection and infrastructure, the evidence of user input in project conception, the project difficulty and the likelihood of success, and the collaboration across collections (http://hul.harvard.edu/ldi/html/current_call.html).

The initial proposal of November 1997 for the creation of the LDI stressed the importance of content, stating that 'digital library developments must be driven by collection content and by relevance to the University's academic mission, not by technology' (http://hul.harvard.edu/ldi/html/ldi_origins.html). In practice, however, content and technological challenges were equally important. Infrastructure development depended upon the selection of a range of formats that would pose different technical challenges. They used the projects to learn from them, letting each content contributor decide for themselves about the materials to be digitized. There was only a general guideline that the material had to be of value to the research community. This might be changing to some extent, however, as in the fourth round of applications the Steering Committee decided to specify areas of

content that they would like to target.

A general observation about selection of material for digitization is that money plays a major role in selection. Grant applicants are designed to find out what is more likely to receive funding, in order to select projects and material with that in mind. In the first round of grants, there was a very open selection process. In the second round, applicants compared with what had already been done and adjusted their proposals accordingly. The condition of the material is not evaluated as part of LDI, since preservation is beyond its funding scope.

The Library would like to build a critical mass of digitized material and has followed a decentralized model about the specific decisions taken. Currently, the Library is focusing more on exploring the issues related to the acquisition of material 'born digital'. Because of the sheer volume of material and the importance of recent 'born digital' material to research, this is a top priority. By comparison, although the digitization of historical collections is important, it is not a key priority, especially since the 'analog' infrastructure for locating the material already exists and has served students and researchers successfully up to now.

Their experience working with different departments of the Harvard community highlighted the importance of honest communication at a very early stage between project staff and the LDI advisors and technical staff. Good knowledge of the material is key. The role and amount of time required for project management was initially underestimated. For example, project management was initially carried out by curators who needed to find time for red-flagging and problem-solving along with other full-time duties.

19.1.2 Overall Obstacles

Some of the obstacles to planning the development of digital deliverables are: technological change and uncertainty; lack of standards; reliance on proprietary solutions for some aspects, but also lack of commercial solutions for others; need for general education around new issues and broad consensus building.

Obstacles to building digital deliverables are: lack of experience in libraries in managing projects in general, and digital projects specifically; the decentralized nature of the institution; the time required to educate the community about new issues and technologies; the greater demands these processes place on the resources (staff and budgetary) of already busy departments; the difficulty of supporting technically decentralized activities (i.e. reliance on many networks, systems, and programs to accomplish tasks makes diagnosing problems difficult and time-consuming).

Another lesson learned is the difficulty of "painting the boat while you are sailing it", that is, building the technology at the same time as trying to fill it.

19.1.3 Selection and Prioritization of Materials

The Grant Review Committee uses various criteria to evaluate grant proposals in the context of all of the submittals in a round. The strengths and weaknesses of each project are considered for each of the areas listed below (ranked as having approximately equal importance) with the exception of intellectual property rights, which must be cleared for the intended use of the material in the project as a prerequisite for funding.

- intellectual property rights;
- teaching and learning potential;
- historical and cultural value;
- research significance;
- improved functionality;
- enhanced access;
- research into digitization strategies.

Also (not in our questionnaire list):

- evidence of user input;
- fit with the program mission;
- evidence of collaboration.

These changed over time, with some criteria being added to the overall program with the addition of strategic areas identified by the Steering Committee. The Committee identified areas that were not already addressed by grant funding. For example in the Round Four Call for Proposals for the Internal Challenge Fund in August 2000 a list of Desiderata was added which mentioned that "LDI has not undertaken work in the following areas yet, and projects that address these challenges may be favored for selection:

- Formats not yet addressed, including: encoded humanities texts, electronic theses; managing access and continued viability of CDs in Harvard's collections, digital video, and web page/site archiving.
- Collecting and preserving digital materials created at the University such as faculty-created materials (ala MIT's D-Space project), formal publications of departments, study centers, etc., and University press materials.
- Service enhancements such as: enriched catalogs (including tables of contents, evaluations by readers, and links to book reviews); campus-wide on-line reference services; and systematic subject portals
- Innovations in scholarly communications - experimentation with changing how scholarly information is stored and distributed." (http://hul.harvard.edu/ldi/html/current_call.html#desiderata).

Three other initiatives are not part of grant funding, but deal with:

- archiving Harvard's digital publications;
- e-journal archiving (Mellon project);
- access project grants (there is separate funding for three rounds of $1 million each for bibliographic projects, improvement of access to traditional sources).

These are all a fairly recent developments. Prioritization criteria did not change substantially over time.

19.1.4 Co-operation

Although the main activities related to the program were carried out in-house, there was wide-ranging co-operation with other organizations including archives, other libraries, museums, academic institutions, corporations, foundations and charities, and historical societies. These spread geographically across the whole range from local to international.

This experience showed that in any collaborative project, it is important to get things right: share standards, agree on best practice, have common guidelines, and colleagues open to communication. All partners need to be very clear when starting a collaborative project or collaborating to build development program.

19.1.5 Purpose of the Program

The purpose of the Digitization *Projects* is to build digital collections of obvious and persistent utility to the Harvard community and, in some cases, the scholarly community at large. Digitization is not undertaken for preservation purposes.

The mission of the Library Digital Initiative (LDI), the digitization *program*, is largely influenced by the challenges to acquire, manage and deliver content that is 'born digital'. It was set up in July 1998 as a five-year program to develop the University's capacity to manage digital information by:

- creating the technical infrastructure to support acquisition, organization, delivery, and archiving of digital library materials;
- providing a team of specialists to advise librarians and others in the University community on key issues in the digital environment;
- providing librarians and staff with experience in a wide range of technologies and digital materials.
- providing librarians and staff with experience in a wide range of technologies and digital materials.

A statement of intent was produced with the Proposal for the Harvard University Library Digital Initiative, which was explicit about the rationale, scope, significance, primary audience, and long-term sustainability of the program (http://hul.harvard.edu/ldi/html/ldi_origins.html).

19.1.6 Nature of the Source Materials and the Impact of Digital Deliverables

The program digitized a very wide range of source material with varied format and nature:

- printed books, serials, and similar bound materials (published) (various sizes up to 11"x13");
- printed documents (gray literature) (various sizes up to 11"x13");
- unbound printed documents (various sizes up to 11"x13");
- typescript documents;
- printed illustrations (up to 5"x7");
- hand-written documents;
- original artworks (Chinese jades, and bronzes, ceramics, various sizes);
- intermediaries (film, microfilm, microfiche, photocopies etc.) (up to 8"x10");
- photographic negatives;

photographic prints (up to 8”x10”);
- sound recordings (cassette and open reel tapes);
- specimens and biomedical images.

The materials selected for digitization were both the entire body of some collections, as well as representative samples, depending on the project. For some projects the digital deliverables were intended to be re-used. For example, the Harvard University Art Museums/Fine Art Library grant project was funded to deliver digital images through an online catalog, but archival versions produced for the project will be used for museum publications. The Baker Library grant project was funded to deliver digital images through an online catalog, but are also used for an online exhibit, as well as a poster and announcement for a library exhibit of the original source material. (description of both grant projects under Round One at: http://hul.harvard.edu/ldi/html/fundedprojects.html.)

19.1.7 Interoperability

For representing content they used GIF for b/w page images and JPEG for grayscale/color page images and pictorial materials. For describing content, they used MARC, EAD, and a localized implementation of the Visual Resources Association Core Categories for many images. The Making of America II DTD was used for brief descriptive metadata and labels association with page-turned objects. The Content Standard for Digital Geospatial Metadata from the Federal Geographical Data Committee and the Draft of ISO19115 are being used to develop a system for finding and using GIS datasets. They use XML for representing structure and the Making of America II DTD for representing the structure of page-turned objects. For sound files, the AIFF format includes structural metadata in the form of channel definitions and time marks.

Because of the decentralized organization of the Library, there are several different systems behind the scenes. There is a need to build a system that works. This is the incentive for individual contributors to make compromises.

For controlling data values they use AAT, Library of Congress subject headings, name authority files, and MeSH, Library of Congress. They use standard vocabularies whenever possible, but they cannot dictate that various projects use the same vocabularies, hence the use of AAT, MeSH, LCSH, etc., and of LC Name Authority File, Union List of Artists Names, and others.

They looked at existing guidelines for digitizing particular document and object types, but did not adopt any (apart from the Ameritech/LC and the MESL ones for projects they participated in). For image and text digitization, rather than prescribe specifications for classes of materials, they developed procedures that create project-specific specifications. The goal is to produce digital reproductions that meet but do not exceed near- and long-term functional requirements (http://hul.harvard.edu/ldi/html/reformatting.html).

In the case of audio digitization, their choice of archival formats relates in part to the systems and expertise resident in Harvard’s Eda Kuhn Music Library Audio Preservation Studio.

In the case of biomedical images, their choice of archival formats (TIFF, grayscale, palette color, RGB, deposited with or without additional Look Up Tables) was influenced by existing practices among scientists and researchers routinely creating digital objects in their labs. The desire is to accommodate practice rather than mandate changes.

They realized that one specification could not work for all users. The technical specifications vary according to source matter.

The primary intended audience for the digital deliverables was described as follows:

“The purpose of the Digitization Projects is to build digital collections of obvious and persistent utility to the Harvard community and, in some cases, the scholarly community at large. … Sponsoring departments for individual projects will be expected to … commit to on-going support to ensure that content made available under the program will continue to be available to the Harvard community …” (http://hul.harvard.edu/ldi/html/current_call.html).

More specifically the targeted audience is: four-year college, graduate school, lifelong, distance, and computer-mediated learning, and archive users.

The vast majority of LDI-related resources are accessible to all, so apart from the targeted audience, anyone who desires to use Harvard’s public catalogs and the linked objects that have not been restricted to the Harvard community, can do so.

They tried to acknowledge the needs of those with disabilities. They took ADA into account in user interface design, and followed locally developed guidelines for compliance with these regulations.

Licensed electronic resources are used a great deal. The new digital content created as part of the LDI Internal Challenge Grant Program has not reached the critical mass that is needed to evaluate its use.

In terms of evaluation of program requirements, generally the Harvard LDI system includes consensus of the community (through working groups, focus groups, open meetings) and oversight by committees.

19.2 Project Management and Planning

Each project has its own project manager. Wendy Marcus Gogel is the project liaison for the central Library for the LDI projects. Project managers are required to provide quarterly updates. The LDI Steering Committee has the program oversight. Other Committees have the oversight for individual systems and/or projects. Project management procedures are project specific and vary considerably. LDI staff have project management expertise and offer advice to individual projects. Templates and guidelines are also available on the web. They always ask in the applications for scheduling, workflow analysis and user needs analysis. Workflow piloting is often carried out, too, mainly on a small scale. They might also request specification of training needs, technical feasibility, workflow analysis, workflow piloting, technology forecasting (whether the project will require new systems or not), and reformatting specifications. Generally staff do not have time for extensive planning. They often want to jump in and do it, but the importance of thorough planning cannot be overemphasized.

Digitization was both carried out in-house (vast majority), purchasing the equipment, as well as being outsourced (35mm microfilm digitization). Preference for in-house digitization was due to the main emphasis of the program on developing local infrastructure and skills. Other important motivating factors were the opportunities for bundled services – materials preparation, digitization, metadata creation (structural and administrative), and loading of data to Harvard's Digital Repository Services. When selecting the outsourcing option, they warn that 'significant time must also be budgeted for other activities: materials preparation and transfer, digital image processing and quality control, structural and administrative metadata creation, file management, and storage' (Chapman & Comstock 2000).

Technologies Used for Image Digitization
Flatbed scanners
Film scanners
Digital cameras
High-end professional cameras

Their experience showed that operator skill is the most important factor in the use of digitization equipment. They established a set of guidelines for document scanning and digital photography. Gray scales, color charts, MTF slanted edge targets, and the Macbeth ColorChecker were used as benchmarks. Some of these were used to generate the ICC profiles. For archival copies the tone and color reproduction was done to the target, not the source file. Detailed descriptions of the Digital Imaging Production Services at the Harvard College Library, specifying the preparation of the facility, staffing, equipment, workflows and specifications are included in the

article by Chapman & Comstock (2000).

DIG photographer Stephanie Mitchell at the digital camera workstation. An array of the technical targets they used to characterize our systems are evident in the foreground (from Chapman & Comstock 2000).

19.3 Human Resources and Training

Apart from the people who work on the various projects, the number of staff who work on the LDI program are as follows:

Type of Staff	Number	% of time on the project
Manger	1	100
Metadata specialist	1	100
Technical development staff	6	100
Projects liaison	1	100
Project staff throughout the University (cataloging, technical processing, programming)	8	100
Reformatting advisor	1	100
Digital acquisitions coordinator	1.5	100

The background and profile of most people on the central LDI team is technical. The remainder have a combination of library background and ICT. Staff were re-deployed from other areas. Although advice was available in-house about technical aspects of digitization, they used external consultants for space planning, studio lighting, and training in the use of specialized benchmarks. An area where training needs have been identified for the main LDI team is project management, where they will organize training in the future. Other areas include application of technical standards, preparation and handling of materials for digitization, post digitization processes (especially color management), and metadata creation. Training was received by the program director, library/cataloging staff, and equipment operators. This was organized in a variety of ways: in-house, using project staff, the Library's own consultants, and external consultants, by attending external courses, with independent study, and by learning on the job. The training did meet the needs of the organization. In-house training is provided to use systems developed by the Harvard University Library. On the job learning and training from external consultants is solicited to solve specific problems/challenges – such as metadata extraction from Kodak Photo-CD images – that arise in specific projects.

19.4 Project Life-Cycle Processes and Procedures

19.4.1 Reproduction and Copyright

Clearance of intellectual property rights is a pre-requisite for funding the projects. Without that, no project is approved. Users can download a variety of digital deliverables. No electronic management systems are used to control copying.

19.4.2 Preservation/Conservation

There is a conservation procedure for original materials. The investigations carried out into the condition of the original materials prior to digitization vary according to project: condition assessment, surface cleaning, re-attachment of photographs to album pages. No conservation activities are undertaken as part of the LDI program, since this is outside its scope. Human and environmental risks were identified during the preparation or digitization process. Human risks include materials packaging and transport, and handling during the process of digitization. Environmental risks include temperature and humidity, dust, lighting, and security. The risk was assessed by the Library Conservation Team which offers advice, training, and treatment. The material does not go by default to the Conservation Team prior to digitization. Risk assessment is also carried out with external consultants as part of space planning in the development of digitization studios. Site visits by the LDI Reformatting Adviser are carried out in consultation with conservators. They use a calibrated light meter to measure lighting intensity and UV. More information is available at the DigiNews article ('Digital Imaging Production Services at the Harvard College Library' RLG DigiNews 4(6), December 15, 2000). Steps taken to minimize the risk include the development of policies for materials transport, security and handling, and upgrade of equipment (e.g. they ordered UV filters for digital camera lighting in one studio). They also use custom-built cradles for photograph albums; they digitize all flat pictorial items under glass at the digital camera in the HCL (Harvard College Library) Digital Imaging Group. The materials were usually prepared by curators and sometimes by preservation staff. Curatorial or preservation staff do not normally prepare/monitor materials during digitization, but photographers/scanning technicians receive materials handling training from conservators in the Weissman Preservation Center and HCL Conservation Services. There are no plans to restrict access to originals once the material has been digitized.

19.4.3 Preparation and Sources

The catalogs used at Harvard before digitization were HOLLIS, VIA and OASIS. The information used from these systems in the digitization process was the unique ID for each item. Access to all the relevant materials (e.g. cataloging information) for the digital deliverables varied according to project. Full cataloging is not required prior to digitization. Unique materials, such as manuscripts and photographic prints, were not altered from the original form for the digitization process. Some items received minor conservation treatment (surface cleaning, reattachment of photographic prints to album pages). Others were rehoused in mylar sleeves with attached labels to ensure that associations between digital files (filenames) and unique IDs for the source items were always correct. Non-unique materials are sometimes photocopied or disbound prior to scanning to facilitate autofeed production of 1-bit page images. No material was rejected before digital imaging. They digitized both originals, as well as using reproductions and intermediaries. For images they used photocopies, 35 mm slides, 4 x 5 or 8x10 transparencies, and microfilm. The form of intermediaries used for audio included 33 rpm discs, audio cassettes, DAT tapes.

19.4.4 Metadata

The originals are generally cataloged. Often the digital materials are not cataloged separately, but are used to represent the original (e.g. for museum objects) or are described as formats, versions, or surrogates of the original (some visual images, page-turned objects). Some of the materials are not cataloged *per se*, but are described as the foundation of the image. In other cases, such as biological type specimens, the original specimen is the cataloged object, and digital images are treated as representations of it. The metadata standards used to catalog originals depend on the kind of material. AACR2/MARC cataloging, archival finding aids, VRA Core records, and purely local schemes are all used in various circumstances. They frequently add a holding for the digital object to a MARC record for the original. However, in some cases it is not necessary to catalog the digital objects as such. While they do generate and maintain technical and administrative metadata about the digital objects, there may be nothing they want to tell the user about the digital object itself other than that it exists as a representation of the original, cataloged object. In these cases there is simply a link from the description of the original to the digital representation of it. OLIVIA uses controlled vocabulary in several fields. The curators decide what thesaurus to use to populate the controlled vocabulary. The metadata records

information about the original object, the digital object, the digitization process, technical details, and administrative information. The metadata records were created by a range of staff, including digitizers, archivists/information professionals, or digitizers who are also archivists/information professionals. Descriptive metadata records for digital deliverables are rarely created as independent entities. To the extent that the digital version is described, that description is available in the same context as the description of the original, whether an AACR/MARC record, an archival finding aid, a VRA-like image record, etc. These different metadata types are served by separate catalogs, but the distinctions between catalogs are not based on the difference between digital and analog deliverables. The catalog for the digital deliverables is available on the internet.

19.5 Format, Resolution and Compression of Digitized Materials

19.5.1 Text

The format chosen for retroconverted text-based digital deliverables was XML markup. They used the Making of America II DTD to encode OCR-generated text that points to corresponding page images (stored as .TIF, delivered as .GIF). The documents did not contain non-Latin characters. The OCR software used was Prime Recognition (system in place at University of Michigan Digital Library Production Services). The level of accuracy achieved was 96% search accuracy (as opposed to character accuracy, which was not counted). They looked at bit-maps, typed in key terms, and checked the system for accuracy. Prior to the OCR process they used TMS Sequioa's ScanFix software to optimize 1-bit page images. The aim of using OCR was automatic indexing. They did not employ any keying in.

The advice they would give to other organizations about to start using OCR would be:

- Produce good-quality 1-bit images. Some sort of enhancement (such as despeckling and deskewing) is advisable.
- Before undertaking scanning and OCR, consider the level of accuracy needed for searching, and encoding needed for structuring the documents.
- Discuss these requirements with the service provider as it may be possible for one vendor to bundle services.

They appreciated the University of Michigan DLPS's service to join words separated by line-break hyphens.

19.5.2 Images

The file formats used for images were GIF and JPEG for delivering (and PDF for page images), and TIFF, Photo-CD and RAW for capturing and preserving. The resolution generally ranges from 300-600dpi for capturing and preserving and 800 pixels (dpi varies) for delivering images. They used 1 to 48 bit-depth for capturing and preserving, and 4 to 24 for delivering. Group 4 and ImagePac PCD compression was used at capturing and delivering stage while JPEG compression, GIF/LZW was used for delivering. The aim of the compression was to reduce cost (by reducing file size, since storage in Harvard's Digital Repository is charged annually by GB), to improve access, to enhance usability, and to decrease storage requirements. Users of the Digital Repository Services are informed about their options (e.g. save money by using compression, but less safe for preservation) and are left to make their own choice. The original scans are retained in uncompressed form, except for 1-bit scanning and Photo-CD scanning projects). Some post-scanning processing is carried out on the images. (For details see Chapman & Comstock 2000.) The average file sizes for grayscale and color images vary from 4-18MB at capturing and preserving stage, and 200 KB at delivery (Museum objects may be scanned at up to 150MB). They use specialized targets and software to measure performance of digital cameras and scanners. These same targets are used to generate ICC (International Color Consortium) profiles for color management.

Their experience showed that collaboration is necessary to arrive at commonly used ways of defining and measuring quality. They do not recommend prescribing digitization methods and technical specifications per class of original material (e.g. color photographic prints).

19.5.3 Sound

The documentation of the methodology and procedures they used to digitize sound is currently in process. The file formats used are AIFF for capturing and preserving and RealAudio for Delivering. They also use Sonic AIFF for capturing. The sampling rate used is 88.2 kHz for capturing and preserving and a variety depending on bandwidth negotiation for delivery. The bit-rate is 24 for capturing and preserving and 16 for delivering. The file sizes created were approximately 30 MB per minute at capturing stage and 1MB/min at delivery.

19.5.4 Moving Image

The program does not digitize moving images.

19.5.5 Quality Control

For the material deposited in the Digital Repository (about 90%) they use both automated and manual procedures checking for the quality of data and metadata. They used a variety of procedures for ensuring the quality of the digital deliverables: spot checks (human); check on a stratified random sample; percentage checks on carriers (e.g. corrupted disks, etc.)(automatic); total check (automatic); and check using MD5 checksums. For metadata recording of the LDI projects it works well when the QA is carried out by the project developers themselves, since they know their material best and can follow the LDI guidelines and keep communicating with the LDI staff. Procedures vary from project to project. For example, a cataloger may have format-specific knowledge, such as how to catalog images, but lack specialized domain knowledge such as Chinese culture or botany . The metadata records and linked images are reviewed by a domain specialist both for metadata content and for accuracy of image links. For some systems, new vocabulary terms are vetted by database maintenance specialists.

Quality control procedures may change the sequence of processing, generate systems requirements for particular projects (e.g. periodic lists of new terms, draft records plus thumbnails), or introduce new levels of review.

19.5.6 Delivery, Access and Use of Digital Deliverables

Harvard provides two levels of access to digital resources: some are open to the public, while others are restricted to the Harvard community.

An important development for the LDI is the Name Resolution Service (NRS), which provides a robust means of identifying and finding digital materials. This is a comprehensive service for creating, maintaining, and resolving names, which are persistent, location-independent identifiers for network-accessible resources. Name resolution is the process of mapping from a given abstract name to an URL that represents a particular instantiation of the named resource. The NRS is comprised of two separate sets of services and servers: (1) an Oracle-based administrative service that manages the metadata necessary to perform naming services; and (2) an HTTP-based resolution service that performs naming resolution (http://hul.harvard.edu/ldi/html/storage_access.html). More information about the infrastructure of the LDI is available on Dale Flecker's article at D-Lib (Flecker 2000).

19.5.7 Dissemination and Publicity

The methods used for informing potential users about the digital deliverables created are project specific with shared responsibility between LDI and the department/unit that is undertaking the project. It may include any of the following methods: announcement on website, press release, articles in print media, print and broadcast media coverage, announcements at conferences and meetings, announcement at electronic listservs, conventional mail shots, or registering with a web search engine. The LDI website includes for example, a page with information about news coverage of the LDI and pointers to the relevant articles http://hul.harvard.edu/ldi/html/ldi_news.html and a page announcing new projects funded http://hul.harvard.edu/ldi/html/funded_projects.html.

LDI can contribute in this area with its vision of being able to "expose" metadata to search engines at large from all types of Harvard sites to assist data harvesting in "deep" web space. They are currently working on the technical development for this. The idea is that users would be able to go to one of the Union catalogs, do a search, and have dynamic websites generated.

19.6 Evaluation, Funding and Long-term Sustainability

No front-end evaluation has been carried out. Formative evaluation was carried out for some projects using paper and online questionnaires, email, and focus group discussions. Some project proposals based their goals on the results of evaluation. For the LDI program as a whole, LDI is committed to a third year evaluation that is part of the original proposal. For individual grant projects, evaluation is required after project completion.

The full five-year program is funded at $12 million from internal Harvard University funds ($7million is for staff, hardware, software, space, overheads and $5 million is the incentive fund). They received everything they asked for.

All elements of the project will need to be updated over time. New materials and metadata will need to be digitized, metadata will need changing, the user interface will need to be changed, and it will be necessary to change file formats. They have a preservation strategy, outlined in the Guide to Preservation Services, that will be made public around March/April. This is the same for all types of material (text, images, audio); for example, file formats will need to be monitored for all media types, storage media and conditions for all media types, and metadata updated for all three categories. The strategy is based on migration of data. Quality control procedures in the life-cycle management are based on monitoring and migration. Longer-term sustainability is not dependent on self-generating funds, but on operational ones.

19.7 Conclusion

As general comments, they mentioned that communication is essential, due to the collaborative nature of the program. Planning and documentation are also very important. Their general advice is: 'once there is agreement, document it'. What Harvard can contribute to the Good Practice Guide (which resembles the research undertaken by the HATII team for the Good Practice Guide), is the recommendation to talk to different players and record the lessons learnt. The effort invested in this shows that mediation and facilitation are crucial and support decision-making. Recipes of how to do things can only take you so far. This highlights the important distinction between getting things done versus getting things right.

Some interesting issues and points arise from this case. One is the decentralized model for content development that Harvard has adopted and how this is adapted in the digital environment: they give a large degree of freedom and flexibility in selection of material to a wide variety of contributors, ensuring rich content. They centralize those areas that nobody else wants to control, such as digital preservation and technical development. They place great emphasis on good infrastructure and architecture of the system. Their experience with different projects highlighted several common pitfalls, such as planning digitization projects based on incomplete inventories of collection, or the general reluctance of people to plan – an area they think is crucial. It is also worth noting that despite their efforts in conversion projects, the Library as a whole is giving a primary role to 'born digital' material. Finally, it is interesting to note that most of the projects that are part of the LDI program are add-on projects for the contributors, and this has subsequent project management and resource implications.

19.8 References

Chapman, Stephen & Comstock, William, 2000. 'Digital Imaging Production Services at the Harvard College Library' *RLG DigiNews* 4(6), December 15, 2000.

Flecker, Dale. 2000. 'Harvard's Library Digital Initiative: Building the First Generation Digital Library Infrastructure' *D-Lib Magazine* 6 (11) November 2000.

valid xhtml 1.1
abp~04/02

20 Indiana University, Cook Music Library

On January 10 2001, HATII interviewed the following members of the Variations project at Indiana University, Bloomington: Jon Dunn, Assistant Director for Technology (Indiana University Digital Library Program); Emma Dederick-Colon, Electronic Music Resources Librarian and Variations Project Coordinator (Cook Music Library); and Michael Fling, Collection Development Librarian (Cook Music Library). The goal of the Variations project is to construct a digital library system to contain the 7000-plus items of audio material that are presently stored in the Cook Music Library. The project also hopes to digitize the objects associated with music, such as scores, texts and videos. The underlying incentive for the project is to create CD sound quality over a broad network, while the project has a subsidiary aim of developing a comprehensive database of scores for distribution.

20.1 Organizational Digitization Program and Policy

The Cook Music Library's Variations project developed from the large amount of assigned listening in courses run by the School of Music. As a result the Library needed to distribute a large amount of music yet maintain control over its playback whilst at the same time improving access to what were frequently limited holdings. Therefore, the Variations project began in 1995 and delivery commenced in 1996 despite hardware and personnel limitations. The focus for the digitization was the assigned listening music, which also suffered from poor playback quality on tapes. The Library's digitization activities were formalized in a policy statement in 1998/9.

The primary objective of this policy is to serve students' needs and service special collections on demand digitization, which has the aim of reducing handling. A secondary priority is to make available faculty performances in digital form. The project has been successful in achieving these objectives and brought faculty on board through demonstrations and the involvement of the Music Library Advisory Committee. Two faculty members also served as "guinea pigs" in 1996. Overall obstacles to planning and building the digital deliverables were faculty's initial resistance to change, technical obstacles such as storage space and CPU time for compression, and copyright and licensing issues.

The primary selection criteria are the materials' teaching and learning potential and enhanced access, as determined by their designation as assigned listening. The Library will endeavor to place any sound recordings on digital reserve so long as they have a copy. For special collections the material's historical and cultural value are also priorities and attempts have also been made to select material to build a standard repertoire, but this has been a secondary priority. Labor cost reduction had been mentioned as a motivation only in selecting material. These selection criteria have not changed over time, but the balance of work has changed as reserve material has reached a plateau. The use of student help for digitization was the most significant change over time.

The Cook Music Library has had informal discussions with other libraries about sharing material and has worked with other academic partners in the UK and Japan (through the NSF).

The primary purpose in creating the digital deliverables is as a teaching and learning resource, in response to previous demand. Public access, depending on copyright and the Library holding a copy, is a second priority; for example, public domain scores are accessible. The project has produced various statements of intent, primarily via internal Indiana University publications, mainly emphasizing the easier access the digital reserves provided.

The type of source material digitized includes:

- printed books, serials and similar bound material (published) – scores;
- printed documents (gray literature) – a trial has been undertaken on CD liner notes;
- sound recordings;
- video recordings – experimental digitization, but the quality was poor with the MPEG 1 format that was used.

The predominant medium of the materials digitized is Compact Disc. The project aims to be a digital library, not just a reserve, and will digitize the entire piece or performance, whatever corresponds to the bibliographic record and not just the section requested. The project had originally discussed re-purposing the digital deliverables but nothing has been undertaken so far.

The following standards, guidelines or tools are used for representing content:

- MPEG 1, layer 2 (delivery);
- WAV (capture).

MARC is used for describing content.

A MARC-based catalog system is used for controlling data values

The project looked for other existing guidelines for digitizing sound materials, in particular for uncompressed WAV files, but they did not exist at the time.

Structural metadata, such as tables of contents in liner notes, are represented in flat text files in a local format. These are then parsed for error checking. The project will probably migrate to XML.

For the project, navigating between ideal and realistic standards is not an issue that arises in relation to WAV files.

The intended targets for the digital deliverables are four-year college and graduate school audiences. As part of a recent NSF grant several faculty members are interested in lifelong, distance and computer mediated learning audiences – for example, to look at a music theory classroom presentation and lesson authoring tool, web based distance music education and the value of a digital music library. The digital deliverables could be used by audiences other than the primary targets; in this case the most likely are researchers and students at Indiana University regional campuses and other universities participating as satellite sites in the NSF project.

The needs of those with disabilities have not been explicitly recognized in the Variations project, but sites that are part of the DLP program have been checked. The digital deliverables in the Variations project are restricted to the music library building because of copyright. Shortly all of the school of music building will be able to access the material. The profile of users for the digital reserve has been as expected, that is undergraduate music students. For the scores material, the audience has also been the same as anticipated, but there have been additional users from outside Indiana University, particularly eastern Europe and South America.

20.2 Project Management and Planning

In-house advice on managing the digitization project was available and the project worked with IBM on technology decisions. The management of the Variations project is entirely integrated into the Library structure, for example the head of the project is the head of the Library and the electronic reserves are part of reserves. There have been some changes in organizational relationships and procedures in that the project is now also part of a larger campus wide Digital Library Program.

Formal project management procedures have taken shape in a recently formed advisory board that wants to involve faculty members. There are no formal managerial quality assurance procedures in place.

A pilot test is being carried out for a project to digitize music scores but was not undertaken for the electronic reserves of the Variations project. In this respect the project staff recognize that Variations is not an example of good practice. The pilot study for the scores project was used for establishing training needs, technical feasibility, workflow analysis, technology forecasting and to assess the benefits to faculty and students. This pilot study is ongoing and it remains to be seen what impact it will have on the design of the project. No formal benchmarking or time and motion studies have been carried out, but the output of each person was examined.

Digitization has been carried out entirely in-house with the majority of equipment bought in (there was some existing audio source equipment). There was not much choice in digitization processes to adopt for audio, and quality decisions were the determining factor. Source equipment ranges from turntables to CD players, and sound cards used for capture are Event Electronics Echo Gina cards. Specialist software used is Sound Forge by Sonic Foundry and Disc to Disk by Microtest.

20.3 Human Resources and Training

A full-time director and project co-ordinator in the Music Library work on the project, but each for a small amount of their time. Slightly over one FT equivalent digitizer is employed (up to 4-5 individuals, 10-20 hours a week each) and portions of a system administrator, programmer and technical manager in the Digital Library Program make up the technical staff at approximately 1.25 FTE. An evaluation specialist is available through the Digital Library Program. These staff all have library, music and/or IT backgrounds and have not been re-deployed from other areas. Where necessary the project looked to other people for advice on the technical aspects of digitization.

Training needs are assessed informally and areas of training that have been identified include:

- preparation and handling of materials for digitization;
- technical operation of digitizing equipment;
- post digitization processes e.g. editing;
- metadata creation;

The team members that have received training include:

- project director;
- specialist technical staff;
- library/cataloging staff;
- equipment operators.

Training has been organized using the following:

- in-house, using project staff;
- learning on the job.

The training has met the needs of the project.

20.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and only owns the copyright in Indiana University recordings, but even this area is unclear. The copyright status of the digital deliverable is declared in a copyright notice recorded by the digitizer (repeating the copyright notice from the sleeve notes). This policy has been effective because students are unable to copy the streaming format. The material in copyright is digitized under the legal provision for libraries (preservation or fair use). Users are unable to copy any material.

For digital audio material, users can only listen to full length compressed versions.

The streaming format acts as an electronic management system as it is extremely difficult, if not impossible, for users to capture or download during playback. As a result the streaming format has proved an effective means of preventing unauthorized use.

The only conservation work carried out as part of the project is the cleaning of originals, such as LPs. Access to the original materials has only been restricted for items in non-circulating special collections.

The cataloging system in place prior to digitization is the Library electronic catalog and a web based reserves list. The project uses all of this information during the digitization process. The project does not have to locate any core reference material to complete the digital deliverables or alter any material from its original form for digitization. Score material will be rejected for digitization if it has blurred pages, is defaced or the print is poor. Nor will the Library digitize any material that it does not already hold or can easily acquire.

The project has digitized from intermediaries, but the material was also available in its original form. Intermediaries used for audio digitization are:

- 33 rpm discs;
- audio cassettes;
- digital sound media (CD and DAT);
- reel to reel.

The original material is cataloged in the Library catalog and the digital deliverables use the same system.

USMARC is used for cataloging the digital deliverables.

Tools used for controlling data values are:

- controlled vocabulary (Library of Congress subject headings);
- thesaurus (AARC2 – music format);
- classification scheme (Library of Congress, de Lerma (for sound recordings)).

Metadata details are recorded about:

- the original object;
- the digital object;
- the digitization process;
- technical details;
- staffing details;
- administrative information.

The existing MARC records for the physical item are used as descriptive metadata for the digitized item. Structural and administrative metadata for the digital item are in a separate set of flat text files, created at time of digitization by the digitizer. The catalog for the digital deliverables is available on the internet and the catalog and object are linked through the MARC 856 field.

20.5 Format, Resolution and Compression of Digitized Materials

The work method used for audio digitization is for one person to digitize and for an automated computer system to save, catalog and link the object.

The WAV file format is used for capturing and preserving audio digital material and MPEG layer 1, level 2 format for delivery. A 44.1 kHz sampling rate and 16-bit stereo bit-rate is used for capturing, preservation and delivery. The delivery compression method used is bit-rate compression at 3.7:1 (from 1.411 Mbps to 384 Kbps) to reduce cost, decrease storage requirements and decrease network bandwidth requirements. The average audio file sizes are 463MB for capture and preservation and 125MB for delivery. No post capture editing is carried out.

Quality control procedures for the digital deliverables are under development. Quality control procedures are in place for the metadata recording and delivery, for example the Reserves Coordinator enters the 856 field and this can be suppressed until the object is ready.

Users have open access to the catalog and the materials are restricted to in-house users. Searching and browsing facilities are the standard keyword and Boolean searches. An encoded sub-set of the metadata can be searched in the same way. The project does not have any special software needs regarding delivery and access. Users are not able to manipulate the digital deliverables apart from the standard playback controls. Users do not have to pay to access the digital deliverables except that all students have to pay a technology fee as part of their tuition fees.

Usage is monitored by automatic data capture and potential users informed of the digital deliverables through articles in print media, conferences, meetings and internal publicity. The most effective of these is internal publicity.

20.6 Evaluation, Funding and Long-term Sustainability

The Variations project has not carried out any front-end, formative or summative evaluations.

The Variations project was largely funded from existing personnel budgets. A new position, Variations Technical Director, was funded 33% by the Cook Music Library, 33% by University Computing Services and 33% by the School of Music. When the Digital Library Program was formed, the function of Variations Technical Director was replaced by parts of three DLP positions: Assistant Director for Technology, Programmer/Analyst, and System Administrator. The University recently received a grant from NSF and NEH through the Digital Library Initiatives Phase 2 (DLI2) program for a new R&D project in digital music libraries, but Variations will still exist as the operational production system. IBM has assisted the project with equipment and other sources have been the Digital Library Program and the School of Music. If the project had had more funding it would have pursued a

more vigorous equipment replacement cycle.

Standards have saved the project money in both the short and long-term. In the short term, using standard file formats, MARC bibliographic description, and a web interface allowed the project to assemble a system more easily and more cost-effectively by picking and choosing products from different vendors that can inter-operate with each other. In the longer term, use of standards should assist the project with preservation issues, since standard file formats tend to be supported by software products longer and have published specifications so that migration is less of a problem.

New material and metadata are added daily with periodic changes to metadata and the user interface. Equipment is updated on a three-year cycle. A written preservation or archiving strategy has not yet been developed, although in general, the project has been carried out with long-term access in mind. This includes use of standard file formats. Migration of MPEG files from current tape storage system to a new university-wide mass storage system has recently been completed, with migration of WAV files currently in progress. The Digital Library Program is working with University Information Technology Services to ensure long-term preservation/access of Library objects in this mass storage system (see http://storage.iu.edu/ for more information on this system and UITS' plan). No quality control procedures have been adopted in the life-cycle management.

The project does not have an exit strategy and should resources not be available sufficient to sustain the digital deliverables, the project staff speculate that they would revert to the previous system of serving reserve material. Loss of the digital deliverables would matter.

20.7 Conclusion

The Variations project at the Cook Music Library, Indiana University, is notable as being one of the few large scale sound digitization projects. The most significant feature was the absence of standards or guidelines for sound digitization and quality control and the adoption of *de facto* standards in file formats, such as WAV, as an alternative. As the Variations project serves only students and faculty of the Indiana Music School it has not faced the copyright problems that would be the most challenging aspect of a similar project that served external users. Indeed, unless an organization owned the copyright in the original materials it seems unlikely that a public service version along the lines of Variations would be possible. The fact that IPR issues are such a potential problem possibly hinders the growth of music digitization and hence the development of standards and guidelines. Nevertheless, the Variations project demonstrates the improvement in access that a networked digital resource can bring and how such technology can be harnessed to improve the Library's service to its core user group.

valid xhtml 1.1
abp~04/02

21 The Church of Jesus Christ of Latter-Day Saints

HATII interviewed Richard Laxman, Herbert J. White and Barry R. Urry from The Church of Jesus Christ of the Latter-Day Saints, on January 29 2001. As part of the Family and Church History Department, the Genealogical Society of Utah is responsible for keeping genealogical records and making them accessible to members of The Church. In order to increase the accessibility of the Department's resources, digital imaging projects were initiated. Digitization was also employed with the intention of preserving the historical documents and artifacts, while also aiding research activities.

21.1 Organizational Digitization Program and Policy

The Family and Church History Department of the Church of Jesus Christ of Latter-day Saints has a massive collection of records which come from archives, libraries, and record custodians from all over the world. The Genealogical Society of Utah (GSU) was founded in 1894 to gather genealogical records and assist members of The Church of Jesus Christ of Latter-day Saints in tracing their family histories. All the records in the collection have one area of commonality, that of genealogy, i.e. birth, marriage, death records and associated information.

Their imaging projects are organized from the first approach and negotiation to the final product and index. The details of this organization are the same for the microfilming projects as they are for digital imaging projects, only the representation format of the documents changes and of course the storage and archiving.

The selection strategy is based on acquiring the records from worldwide sources for the Family and Church History Department. The main associated drivers for the imaging projects include: Historical and cultural value, research significance, improved functionality, enhanced access, research into digitization strategies, preservation (e.g. against content loss), potential to reach disadvantaged groups across the globe.

This will not be altered over time, as the organization has one driver for collecting the information and that will not change. The projects they work with have varied drivers for collections and selection of material but the GSU has one aim in all imaging projects.

The very nature of the imaging projects means that they collaborate regularly with a variety of organizations including archives, record offices, libraries, and religious organizations. This collaboration is at all levels. A more detailed description of how this collaboration is achieved is given in Project Management and Planning.

21.2 Management and Planning

Each imaging project undertaken by the GSU follows these six stages:

- Negotiation and project administration

This stage covers all from authenticating the images (to ensure that the data is appropriate), negotiations with the archives to achieve permission to acquire and to enhance the material (where increased readability is required). This raises the issue of copyright when the images are enhanced. This is agreed at the time of negotiation. The Family History Department negotiates the rights to reproduce, distribute, display, use and permit the use of the Digital Images, for nonprofit purposes, via any technology and in any medium. The goals at this stage are to define a win - win situation for all parties (record Custodian, GSU, and patron), to provide better access and to establish good practice in managing imaging projects.

- Capture Convert Acquire

This stage is the technical process of capturing the digital images.

The subsequent stages of the project are described in the following subsections:

- Image and metadata processing: 21.4
- Storage and preservation: 21.5
- Indexing and cataloging: 21.6
- Access and distribution: 21.7

The Six Stages have been designed by the in-house project management staff and is in place after many years of microfilming projects.

The Six Stages cover the process from the Records Custodian either approaching or being approached by the GSU to digitize the records, to the digital images being indexed and made available to patrons. All the processes are broken down to discrete units that are managed by the appropriate team from the GSU, e.g. digital imaging by the Field Services and Support Division and volunteer indexing by the Extraction Section.

The source material will vary widely from project to project and can range from handwritten documents to typescript text. The content is always consistent with the remit of the GSU.

Standards are vitally important to the GSU. Not only in adhering to accepted standards, but also in helping to establish new standards in the digital imaging arena. Staff are members of AIIM - Association of Information and Image Management. AIIM is the standards developer in the United States responsible for writing standards for the Document Imaging Industry. GSU staff, working in conjunction with AIIM and the other volunteers who have interest in Document Imaging, are in the process of identifying the areas where standards are required in this Industry. Once written, the standards are then forwarded to ANSI to be adopted as an American Standard, and to ISO via International Standards Technical Committee 171 (TC171) where the standard will be changed to reflect the needs of the international community and adopted as an International Standard. An example of such a standard is the recommended practice, which the GSU is writing regarding Metadata for document imaging. This document will include the Categories, Elements and Sub-Elements needed to describe a document, which is being digitized and will be made available to a general audience. This document will include an XML DTD, which will be mapped to EAD, Dublin Core, and other defacto standards. They use standards for improving interoperability within projects and also to further the research and development of preservation materials.

As the GSU has built on the experience of the microfilming projects (that are still currently happening), they have a whole organization behind them with the support required to ensure that the digital program is successful as the GSU is an integral part of the larger organization of the Church of Jesus Christ of Latter-day Saints.

21.3 Human Resources and Training

The GSU has a large staff of specialized people working in different areas of the digital program. These core staff are supplemented by paid employees and teams of volunteers that work in a variety of areas in the digital imaging projects. These volunteers come from the Church as missionaries. Typically, they are retired couples who are assigned to the imaging projects. They can come from a variety of areas not necessarily linked to digital imaging but where possible, skills are targeted to appropriate areas. The process to becoming a missionary is through established procedures in the Church before being assigned to a specific area. It can be a two-way approach.

Volunteers are then given a two-week training program in digital imaging, which can then be applied to any project regardless of the type of material. A volunteer will spend 12 - 24 months on a project. They fund themselves in accommodation, food and everyday expenses, but GSU may help with travel where necessary.

This training is intensive and covers four main areas: document preparation, imaging, quality assurance of images and indexing. The main training is in using the digital camera that will be used in the projects. Any on-site training required is met by the on-site project manager. There is a regional project manager and an area manager.

The training schedule takes the volunteers from basic digital concepts and computer basics through to setting up the camera from scratch. The training manual is comprehensive and enables the volunteer to refer to it when on site. The GSU have found it simpler to train volunteers to capture images digitally than to use microfilm. The digital operator is now the evaluator, thus removing one link from the chain. They try to match volunteer's skills with the skills required for the project, such as document preparation or imaging. The issues with the archive and how to deal with the materials are dealt with on-site. The GSU sends out the volunteers with training on the

practical use of the digital camera and software and the associated software and the skills are cemented on-site.

The volunteer management is part of the larger organization that drives the digital imaging projects. Having project managers at various levels throughout the world enables the volunteers to adapt to the project after the initial training as support is available at all levels. Technical support is through the nearest technical staff, i.e. Germany for Europe. At the moment, there is one digital project in Scotland. However, the basic structure was in place from the microfilming process and it is the GSU desire to move to digitization as the future of imaging that has adapted the process for digital imaging training.

The success of the volunteers is based in the organization of the whole and the levels of management that are established from the start. Rarely do volunteers have to stop working on a project as the training ensures that they are able to work on all the processes in the projects.

They are constantly updating the training to meet the needs of the volunteers and to adapt to the changing technology.

21.4 Project Life-Cycle Processes and Procedures

The copyright issues can be a major factor in the imaging projects. These issues are identified and agreed at the negotiation stage of the process. The GSU very rarely holds the copyright to the materials. The negotiation process can be lengthy, in order to achieve the goals required by the archive, the GSU and ultimately the user. These delicate negotiations are a vital part of the whole process. The Family and Church History Department has to eventually make the materials available to genealogists across the world, so the rights to distribute the information must be attained.

Users will be able to view and possibly print and download a variety of objects, from straight indexes to JPG images of the digital material. There may be minimal charge made for this to cover costs.

The archivist on-site generally prepares the material and the volunteers who do the imaging have basic training in handling delicate materials. The material they work with will have been prepared by the curators where necessary. The GSU is occasionally involved, at the request of an archivist, in the preparation process, but usually only handles the objects in the digitization area.

The GSU has a digitization system in place for their external and internal projects. This is the capture, convertion or acquisition stage. In digital projects, they are capturing the image from original materials. They have a system that includes the digital camera, the software to acquire the image and create metadata automatically that links to the main catalog and index to assist in retrieving the digital image and information from these systems. The imaging process also includes a simple and efficient method of quality assurance of the images.

The GSU has recently acquired another brand of digital camera, which has an 8 megapixel range. The camera set-up has daylight balanced fluorescent lighting for cool lights and baffles to protect the eyes of the camera operator. Each computer attached to a camera has three hard drives. One is used for the camera software. The other two are configured in a RAID format to rapidly store the images at capture time.

The camera has a three-pass RGB system to capture color images. Their software is specially written for the camera and for archival projects. They use color and grayscale targets to calibrate the camera, typically they will do a large-scale calibration once a week to ensure the LCD filter has not shifted and color charts at the start of day. The lights have been an issue. They typically last 11, 000 hours and can add to color shift but the large scale weekly calibration ensures this is kept stable. Recent enhancements to the software provided automated color balance. Exposure is checked daily. The capture and save process takes 6 - 8 seconds for color images, 1 – 2 seconds for grayscale and bitonal images. If the image is deficient in any way, it is overwritten. The software allows the camera operator to recapture the image, automatically write over the unacceptable image, keep the image file name the same, and ensure the log file is not in error.

A log file to hold a limited amount of metadata is created automatically during capture. This file will hold up to eight or more elements to be defined by the operator and could include: the operator’s name, the filename, the camera details, etc. The file is automatically updated at each capture. The operator will have to occasionally enter the name of the next large segment of material, for example, a new register book.

The Family and Church History department has begun to develop metadata categories for the digital images. This development is crucial to the longevity of the images. They are researching various categories of metadata to include: discovery metadata, preservation metadata, rights and security metadata and representation

information metadata. They have evaluated the Dublin Core as well as many other systems but are favoring Dublin Core for the discovery and description category. The breakdown of their metadata includes:

- ADMINISTRATIVE; rights and security
- DISCOVERY AND DESCRIPTION; Dublin Core for this area
- Extraction for GedCom; age, date, genders, names, index data, occupations,etc
- PRESERVATION
- TECHNICAL and capture
- USE

Each metadata element is defined using a set of attributes. Some of the definitions come from the ISO/IEC 11179 standard for the description of data elements. Others come from the Family and Church History Department.

The metadata is the key to the index and catalog that enables users to conduct genealogical searches. The GSU has an extraction program that is staffed by volunteers as the capture programs are. These volunteers extract the genealogical data and indexes from the digital images.

The catalog to hold the extracted information has been developed specifically for the GSU by specialized developers. They are also working closely with Brigham Young University (BYU) to develop sophisticated techniques to assist with the extraction. This can be extremely complex due to the nature of the images, i.e. hand writing and language. The catalog will eventually also be linked to the images for access.

21.5 Format Resolution and Compression of Digitized Materials

As explained above the capture stage is controlled by automating the process as far as possible and by using the same system for all capture projects as well as by constant research into the latest technology for digitization.

The images produced from the capture process are TIFF files.In one project they produce three images: one full color uncompressed TIFF for archive, one full color JPEG for the patron and one grayscale image for the GSU. The JPEG is for access purposes. The capture format standard for the GSU is uncompressed grayscale, since all of the information on a given document is legible on a grayscale image and the file size is smaller. The image files are linked to the metadata file, which is created during the capture process. Typically a color file will be 14 - 18 MBs uncompressed, grayscale 6MBs uncompressed, and Bitonal 150 - 250 KB compressed. The patron will have a compressed image of sufficient quality required for their purposes and the GSU have their images in grayscale or bitonal where the quality is adequate.

To determine the correct resolution, the GSU has established a standard for the capture of the characters on the document. The process is known as Pixels per Line Segment or PPLS. Prior to capture, the operator will find the thinnest, lowest contrast character represented within the documents to be captured. Tests will be made on the line stroke of the character to obtain 3 - 4 pixels in the grayscale format and 2 - 3 pixels in the bitonal format. Through this process the operator can be assured that all characters with line segments at that level or above will be captured and made readable.

Some images, which are captured, require image enhancement by virtue of the type of documents that are being acquired, such as documents that are 100 years old or older and have aging blemishes. Some images are enhanced, as the nature of the material generally requires enhancement to make the data extractable. It is in this area that the GSU is working closely with BYU to develop algorithms to enhance the images. They are using technology to clean up fuzzy typescript, to de-fog and remove image distortion. Interestingly, they are also researching the possibility of OCR for handwriting. The computer scientists at BYU assisted in the development of the lasso tool for Adobe PhotoShop. It is technology such as this that is being developed to enhance old handwriting of documents with a large noise factor. The goal of the Family and Church History Department would be to produce these types of technology and then make them available as public domain for other Archives and Libraries to use, to make their documents readable. See http://python.cs.byu/gendocs/gendocs.html for more information on the algorithms used for the image enhancement.

Quality assurance of the images is an integrated part of the capture system and uses software developed for the programs. The camera operator will examine each image and reject for skew, readability and color and lighting balance. If rejected, the image will be recaptured. The defective image will be written over and noted in the logfile or indexing.

When the images are sent to the GSU from the projects, there is an audit program that further ensures quality. This audit program is once again specially developed for the process. A wizard sets up the rejection threshold based on the number of images in the sample and uses a statistical sample created by a random number generator. An auditor checks each random image. There are currently 24 reject criteria - these criteria are identified by codes. Once an image is rejected the code(s) will be associated with image file name as part of the image log file. Based on the number of images in the process and the rejection thresholds which have been established, if the maximum rejection number is reached, the audit program turns off and the whole batch must be re-imaged. Auditors are trained to use this system and to evaluate the images. Typically, they look at 150 images in a batch at their choice of speed, e.g. one second per image. They are evaluating future auditing processes that will use histogramic analysis as well as checksum to automate the audit process.

21.6 Evaluation, Funding and Long-term Sustainability

Evaluations with users are handled by a different area of the Family History Department and are ongoing as the technology develops.

Preservation is crucial to the GSU, their data and information must be available for as long as possible and as accessible as possible. Currently, they use DLT tape and spinning discs for the storage of the digital materials. The tape solution is used, as this enables the material to come directly from the field. Backup tapes are stored off-site in the Granite Mountain Records Vaults - a secure, climate controlled underground storage facility in Utah. The index is the key to the tape access. Once again, they are working closely with BYU to explore digital storage and long term preservation.

Norsam (http://www.norsam.com) is the current method that seems to have the most promise for long term storage of digital materials. This technology etches the information on the disc using ion beams, and is technically visible to the human eye, with the use of a powerful microscope. These discs can hold up to 17,000 images etched on a two-inch square plate and 165 GB data compared to 4.7GB DVD or 0.6GB for CD.

It is not only the technology that ensures the longevity of the materials, it is the migration of hardware, software, media and people skills and the GSU is constantly reviewing its management procedures to ensure that no information can become obsolete due to hardware or software or media or even skills being lost.

The work with AIIM and ISO and ANSI is assisting these goals in the creation of GEDCOM and XML DTD extracted from the EAD written to tackle the issues of digital data capture and information retrieval. The extraction program and index will rely on this as will the metadata at the capture stage.

Access to the digital materials is covered in three areas, the index to the images, the catalog and research guidance. The index is linked to the catalog which in turn will lead to the digital versions of the records. The GSU plans to have various levels of access; worldwide access, the internet and intranets, through libraries and homes, through electronic media distribution and also through human readable media distribution. The goals are to have what you want, when you want it, where you want it, in whatever format you want it. The main drivers are improved readability through their work on image enhancement, viewability and accessibility. At the moment any charges cover cost of distribution. Access to the materials on line may or may not remove this cost. However, if users want hard copies or copies on CD for example then a cost recovery charge will be made. It is the nature of the materials and the nature of the digital program that drives this large-scale accessibility.

valid xhtml 1.1
abp~04/02

23 Museums and the Online Archive of California

On January 25 2001, HATII interviewed Richard Rinehart, Director of Digital Media and Guenter Waibel, Digital Media Developer, from the Museums and Online Archive of California. The Berkeley Art Museums and Pacific Film Archive is one of a number of organizations that undertakes digitization projects, to then have that material stored by MOAC. The main incentives for digitization are issues of distribution and extended access to the unique materials contained in various institutions throughout California.

23.1 Organizational Digitization Program and Policy

This report discusses the Museums and Online Archive of California with particular reference to the Berkeley Art Museum and Pacific Film Archive. BAMPFA is the driving force behind MOAC due to the involvement of the Director of Digital media at BAMPFA as general MOAC project manager.

The collection at Berkeley Art Museum and Pacific Film Archive was already cataloged and MOAC chose materials to digitize from sub collections. The main priorities for choosing the materials are access, logistics and research. This is a three way process with the digitization team, the curators from the museum and the collection managers deciding which collection will be digitized. Aspects such as copyright, accessibility and research demand are taken into account, as are physical properties for digesting. This process is not formalized.

An example of this process is provided by the Cha collection, currently being digitized. The curator recognized this collection as important and distinctive material for research. The collection is mainly on small paper and photographs so there is no difficulty with physically moving it around the museum, a concern of the collections manager. And, as such, it was simple to digitize as there were no intermediaries and it could be done on the flatbed scanner. This sub-collection therefore met the requirements of all three parties in the museum and is now being digitized. More technically complex materials, such as Chinese scrolls, await technical developments before they can begin to digitize.

Project management is simpler with sub-collections as they are a discrete section and often funding can be targeted at sub-grants for specific collections.

The selection criteria include historical and cultural value, research significance and preservation against content loss. The criteria will change depending on the variables of collection management and new technologies.

MOAC is a consortium project and thus collaborates with other museums and libraries listed below to ensure that the digital versions are accessible through the online database. Different projects will have different levels of co-operation, including local, national, regional and international. Partners include the Tate and the National Gallery of Canada in a conceptual art project. Collaboration can be difficult to establish but is well worthwhile in terms of scope and funding. Collaboration also assists with technical expertise across the various projects and institutions involved. MOAC enables smaller projects to act independently and make their own decisions while still being part of the larger project. This modularized view enables the projects to be self-supporting, while following the technical requirements for MOAC using the database tool that has been specially designed for it.

The institutions involved in MOAC are:

- Berkeley Art Museum / Pacific Film Archive;
- Japanese American National Museum;
- Oakland Museum of CA;
- Phoebe Hearst Museum of Anthropology;
- Iris & B. Gerald Cantor Center for Visual Arts at Stanford University;
- UC Berkeley Bancroft Library;
- UC Berkeley Museum of Paleontology;
- UCLA Grunwald Center for the Graphic Arts;
- UCLA Fowler Museum of Cultural History;

- UCR/California Museum of Photography;
- California Digital Library.

Digitization has been ongoing within BAMFA since 1994, with the main purposes being research, preservation, public and widening access, and teaching and learning. MOAC has a web page offering project descriptions as well as technical standards and papers on imaging and tools.

The type of materials varies from project to project and is typical of museum collections – artifacts, paper based documents, original artworks, photographs, etc. Due to copyright issues, the nature of BAMPFA means that while some film is digitized, it mainly consists of video of artists discussing works, rather than actual creative films,

MOAC uses standards where possible for representing and describing content: CIMI, JPEG, MPEG, SGML, XML, MARC, Dublin core, EAD, CDWA. EAD is the main standard used and the database they have designed enables the data to be output in EAD format for use on the online systems. This approach allows sub-projects to be independent of each other whilst still adhering to the technical specifications for metadata and finding aids. The virtual museum that MOAC represents can function because of these standards and create a standard access area, with diverse collections from museums in disparate geographical locations.

Standards for data values include AAT and LC subject headings.

MOAC believes that standards should be used from the beginning and as much as possible though museums should be able to adopt as much as is practical for their purpose. The standards ensure long- term legacy data while in the short-term enabling projects to create the resources. The reason for the use of many different standards is that there are theoretically ideal standards but no commonly available tool that is affordable and functional for all projects; hence the creation of the MOAC database.

23.2 Project Management and Planning

The digital media department in BAMPFA is part of BAMPFA's administration department. The department is now closely aligned to the photography and collections departments and is integrated into the overall structure of the museum through this collaboration.

The museum is in the process of constructing a whole new building for the collections and a digital imaging lab is being planned from the start, as well as workstations in the exhibition rooms. This foresight is due to the collaboration and integral role the department has created for itself.

There is a sub-board of museum trustees who have a responsibility for the digital media department and assist in the support and advancement of the work it does.

MOAC's project management is intensive since it co-ordinates the digitization of resources in eleven museums. It uses project management software to assist in this task. The funder does not impose project management structures, so the level of management often depends on the project. BAMPFA is currently involved in four projects, including MOAC. These projects may be led from the front with advice but the project management is fully self-determined.

As with many projects, the actual digitization is both carried out in-house and out-sourced. BAMPFA outsources materials where capability is an issue, for example it out-sources 16mm film to Mini DVD. The current director bought the in-house resources when the digital media department was established. They have flatbed scanners, slide scanners and a BetterLight high-end digital camera. Very detailed specifications are employed for data capture on format and resolution, for example. MOAC partners use grayscales and color charts for data capture.

23.3 Human Resources and Training

The digital media center in BAMPFA has four staff – one director, one digitizer, one technical support worker and one webmaster. As is common in these projects, the staff came from a variety of humanities backgrounds with an IT interest and associated skills.

Training is important, as the center relies on current staff for a large output and staff must be able to respond to new technology advances. They now have a budget that enables everyone to attend at least one conference a

year, in order to appreciate new developments.

Training is carried out on site and through learning on the job as well as courses, internal and external.

23.4 Project Life-Cycle Processes and Procedures

The collections in BAMPFA are partly in copyright and partly owned by the museum. The catalog does contain this information. Material in copyright is digitized but may not be mounted online for public access. Due to the contemporary nature of the museum collections, much of the material that is not owned by the museum is the copyright of artists who are still living or have a sympathetic estate and often grant permission. Material that is digitized but where there is no owner's agreement is created under license to the copyright holder. However, the copyright situation is a criterion for selection of material in the first instance.

BAMPFA and MOAC's preservation strategy has been developed from the experiences of other partners in the Californian Digital Library. They use fixed media CD as storage and are exploring DVD. The aim is to replace and migrate to ensure longevity. The CDs are physically stored in the building, two copies for redundancy. They eventually hope to store both on and off-site, particularly because of the geographical location of the museum. The database enables access to the CDs by an index number. The backup run by UC Berkeley automatically supports the database, meaning that they do rely on campus support. The access materials on Sunsite are stored in servers online but they are in JPEG low-resolution format and not appropriate for long term preservation. This is established by storing the images on CD as uncompressed TIFF. They are looking at QuickTime, DV, and MPEG7 for audio and video as a possible format for storage. The metadata on the CD is minimal; there is a list of names and files, and the catalog must be consulted to match the file with the metadata. The file name is a type of metadata, in that there is an institution acronym and a physical collection number. However, this is a point of contention and may be revisited.

The Digital Assets Management Database (DAMD) was developed by the digital media developer and is based on File Maker Pro. The database allows the cataloging of the digital materials with links to the main catalog that the museum has for its collections. The database was designed with museum needs in mind, so the structure and content holders (i.e. the records and fields of a database) allow easy cataloging of museum material. The main aspect of the database is that it enables the data to be output (at the click of one button) to EAD and MOA2 XML files, in accordance with MOAC specifications. The most important aspect is that it is platform-independent and is flexible enough to be used by all MOAC partners, regardless of specific project features. The EAD XML file acts as the finding aid to the images and is in accordance with the finding aids developed for the Californian Digital Library.

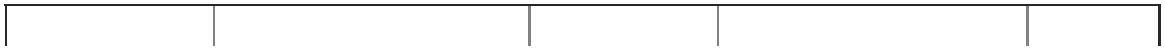
The metadata attached to the images ensure that the image catalog and the main catalog can communicate. There is a controlled vocabulary within DAMD, but it was created in-house. They do not record information about the original object, just the link to the main catalog. Both catalogs are written in File Maker and the link is in the relationship. They record information about the digital object, the digitization process, technical details, staffing details and administrative information. The metadata is created by the digitizer in the center, and the metadata elements are approved by the MOAC committee for the MOAC partners to use with the DAMD. In turn, these elements are a subset of the CDL specifications:

"There are currently 20 required metadata elements, two of which need to be manually input for each object. The other elements would normally be automatically generated or inherited from 'default' fields set in digitization management software."[6]

It is the specification of these standards as subsets of overarching specifications that ensures that the material produced by individual museums and MOAC are interoperable and readily available for the CDL access system.

23.5 Format Resolution and Compression of Digitized Materials

MOAC has specifications for image capture that all partners adhere to, which are once again based on specifications from the CDL. The following table is taken from http://www.bampfa.berkeley.edu/moac/moacfullspecs.html#ImageSpecs



	Resolution	***Format***	***Bit-Depth***	***Gamma***
Master Image	3000 pixels at least longest edge(no matter source size)	TIFF (no compression)	24 bit color for color sources 8 bit grayscale for b/w sources	2.2
Intermediaries	Anything over 150 pixels longest edge	Optional. JPEG, QTVR, Flashpix	24 bit color for color sources 8 bit grayscale for b/w sources	2.2
Thumbnail Image	150 pixels exactly longest edge	JPEG (JFIF)	24 bit color for color sources 8 bit grayscale for b/w sources	2.2

These specifications allow all projects to meet the minimum criteria set out by CDL and MOAC and ensure the interoperability of the materials.

Images are stored as TIFF and displayed as JPEG in thumbnail and larger format. The aim of the compression is to improve access and decrease storage, and while not a major influence, theft was an argument used for compression. A little image processing is done after scan time, cropping, resizing and unsharp mask in PhotoShop.

As discussed before, BAMPFA is digitizing video, mainly films of artists talking about their work, for long-term research and not merely for temporary exhibits. The digital video objects are all held in the database as items identified. They use DV/QuickTime for capture and are exploring this format as well as MPEG7 for storage.

Almost no quality control is carried out. There are procedures on creating the images and as the files are saved with the occasional re-check of the images. However, for the most part, the digitizer carries out quality control at the time of scanning.

Access to the materials is at various levels: internally through the DAMD and externally through the BAMPFA's own webserver, and the Online Archive of California central server using DynaWeb. The EAD XML format enables MOAC and its partners to format their material so that it can be mounted immediately on the Online Archive of California central server. DynaWeb makes SGML files HTML compatible, and which can thus be viewed on the web. The advent of XML and the possibilities for web delivery are being explored but DynaWeb is currently the preferred access option. The finding aids produced from DAMD provide the key to accessing the digital objects, as do the specifications set down by the various levels of the CDL.

23.6 Evaluation, Funding and Long-term Sustainability

The digital media center at BAMPFA has an annual budget of $190,000 but this includes the technical operations budget, i.e. providing IT facilities for the whole museum. Special Project funding comes mainly from US federal funding, and one staff member is on “soft money”. They have had endowments and corporate sponsorship. Without a doubt, using standards has saved money and ensured the success of the project – not just community standards, but the specifications set out by CDL and MOAC.

23.7 Comments and Conclusions

MOAC represents a new direction in collaborative projects, having as its objective every museum in the state

sharing their resources and collections. The broad thinking behind this has led to the creation of guidelines, the adoption of standards and the creation of tools to enable the vision to succeed. As part of the OAC, they have experience and understanding of the practices and processes required to create and develop a digital imaging project.

The database purpose-built to output EAD is significant as it highlights the problems associated with moving the images from the internal catalog to web based access. Prior to this database, the information had to be extracted, often by hand, to establish the data required for the EAD finding aids. Such tools are crucial for collaborative projects to ensure that the relevant material is recorded and extracted.

The guidelines that MOAC has developed also assist in ensuring that every museum which becomes a partner has the minimum standards required to create the quality of digital images that MOAC prescribes.

While the mechanisms are in place to facilitate the recording of metadata to ensure the longevity of images, the quality control and storage strategy are not formalized, thus leaving the possibility of errors and disasters. Staffing is also a concern. MOAC is managing with a small core staff but larger resources would enable formalized quality control thereby ensuring the longevity of quality images.

---

[6] http://www.bampfa.berkeley.edu/moac/moacfullspecs.html#ImageMetadata

valid xhtml 1.1
abp~04/02

24 New York Public Library (NYPL)

Barbara Taranto, Director of the Digital Library Program at the New York Public Library, was interviewed by HATII on November 26 2001.[7] The DLP is responsible for overseeing collection surveys of the NYPL's research libraries. The Science, Business and Industry Library is one such department whose resources are being considered for inclusion in the digitization process. The materials are assessed in terms of their historical and cultural value, as well as their research significance. The Program hopes to allow a more inclusive audience to have access to the materials held by the research libraries, while broadening the overall vision of the NYPL.

24.1 Organizational Digitization Program and Policy

Over the past year, NYPL has developed its digitization strategy from having no system in place, to the creation of a robust digital library program suitable for many different materials. The impetus behind the establishment of this more formal strategy was $5 million worth of funding. This was originally in place to digitize 600,000 images over a period of five years, but it has now formed the foundation for the wide-ranging Program. There was an initial collection survey of sorts undertaken that prompted the original approach to donors. Now there is a more systematic survey to identify materials.

The NYPL is in two parts, the public branch libraries and the research libraries. The collection survey to identify material has been completed across the research libraries, of which the main ones are SIBL, Lincoln and the Schomburg center for studies in black culture. This survey is managed by the Digital Library Program (DLP) and is advised upon by a committee of curators, appointed by the library. The primary objectives are ease of access, ease of use, ease of navigation, and as quick a delivery as possible. A number of lessons have been learned from the first survey such as the discovery of scattered collections, problems of absent metadata or metadata divorced from the collection, the ease, or otherwise, of scheduling scanning, and curators putting forward material for their own reasons. Collections on arrival to the library had been broken up and given to different locations within the library (depending upon, amongst other things, the rarity of the objects and the content of the materials). One of the major aims of the Program is to re-unite these scattered materials and create an intellectual intellectual landscape with complete collections. Similarly, the technical metadata now specifically refers to the objects, whereas in many cases there had been no associated technical metadata.

The collection survey was used to establish a preliminary priority of how much material could be contracted out (the standard material). The project also looked at other US digitization projects. The Deputy Director for Research Libraries established these priorities in conjunction with curators. These priorities were then formalized by the DLP in a strategic policy statement, which was approved by the DLP and the advisory group.

Rather than having production as a main objective, the library is concentrating on developing an infrastructure for archiving as a model for standardization and creating an OAIS model for ingest format. They are working towards an open source system with attached metadata, but which is not embedded in the actual object. They have taken this model from the Harvard system. For each session of creation all the metadata is bundled in XML with the object. The program is also working towards the creation of multiple search options designed around the varied needs of user. Once complete this will offer searching by text entry, and visual entry.

Broadening access, social inclusion and reinforcing the NYPL's presence in the new library forum are subsequent objectives. The project has not yet achieved these objectives because the funding came through more quickly than the original project envisaged, recruitment of qualified staff has been difficult and production schedules slower. A recommendation for organizations attempting to formalize their selection criteria in strategic policy statements is to adopt formal participation procedures.

The one overall obstacle to planning the development of the digital deliverables has been recruitment, and for building the digital deliverables the timing and establishing of the proportion of in-house and external work.

The criteria that have guided the project in selecting and prioritizing materials for digitization are as follows.

Selection is restricted to pre-1923 material. Historical and cultural value is key. Research significance is important, as are enhanced access and social inclusion to reinforce the Library's public reputation. Preservation against content loss is also a significant factor. The material's teaching and learning potential is less of an issue and, although conservation is important for curators, it is a middle priority for the project, along with improved functionality and potential commercial exploitation (the latter only through reproduction). Publicity is important in the general sense and research into digitization strategies will become important from year three. Space rationalization and labor cost reductions have the least priority and will only be a factor through reduced handling. It is too early to say whether these criteria will change over time.

The project has not co-operated to date with other organizations in developing its digitization program. The current status of the project is ongoing. The DLP is indefinite and its establishment as a fifth research center/library of the NYPL is possible. The start date of this phase of the project is 2000 and anticipated end date is 2004.

The purpose in creating the digital deliverables is research for the NYPL and profession at large. The project has not made an explicit statement of intent.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- unbound printed documents;
- typescript documents;
- printed illustrations;
- hand-written documents;
- artifacts;
- original artworks;
- intermediaries (not microfilm, since they have this already);
- hand copies;
- photographs;
- film and sound recordings (some pre-1920, currently on tape, to be included in a DVD archive).

The nature and format of the materials digitized is wide and varied; photographic negatives and postcard-sized prints are outsourced. Incunabula had been looked at for inclusion in the Program. This was in reponse to a research project in collaboration with Princeton University studying their typeface with mathematical models. Digital microfilm already exixts at the library, but the project is working on a 'page-turning' program for it. Projects do sometimes lead the Program into new avenues; these can originate within, or outwith the Library. The digital deliverables represent neither a sample nor an entire body of material — the aim is to create an entirely new collection. At present there are no plans for the project to re-purpose the digital deliverables, but there are some ideas.

The following standards, guidelines or tools are used for representing content:

- TEI;
- Unicode (will be using);
- JPEG;
- METS (a new frame-based ingest format);
- SGML (now rarely used, and it is hoped all SGML will be converted to XML);
- XML.

The following standards, guidelines or tools are used for describing content:

- transcription to MARC (as master record);
- Dublin Core (to a lesser degree);
- XML and clearly defined DTDs.

The Program is developing a system which involves traditional hierarchies of content description, but also allows resemblancies across these hierarchies to be displayed, which is rarely available. To do this they are using a MARC system for hierarchies and a relational database to chart resemblancies. This is in progress.

The following standards, guidelines or tools are used for controlling data values:

- LC Subject Headings;

LC Subject Thesaurus (both in database).

The program looked at other existing guidelines for digitizing particular document types when planning its digitization strategy, including LC and Berkeley.

The following standards, guidelines or tools are used for representing structure:

- DC headers in database.

The recommendation in relation to standards in general and for navigating between the ideal and the realistic is to adopt some general standards but not necessarily all the details in practice.

The primary intended audiences for digital deliverables are graduate school, lifelong learning, distance learning and public library users. Secondary target audiences are K-12, community college and four-year college. The project has not carried out any evaluation of the target audiences but will do so when a research post is established.

The deliverables can be used by audiences other than the primary intended target audience. The project has acknowledged the needs of those with disabilities via the W3C's "Guidelines for Web Site Accessibility" but is not sure to what extent. No limitations on the use of the digital deliverables are imposed directly. However, to protect the Library's rights in the digital deliverables the quality is limited for the freely available material, which prevents it being re-used for publishing. The Library is also protective of its reputation and name and wants both to safeguard possible income and not misuse sensitive material.

24.2 Project Management and Planning

There has been no external project management and one member of staff has a project management background. The management of the project is part of the federal structure of the NYPL, means it can take longer than desired to reach decisions. There have not been any changes or re-evaluations of organizational relationships or procedures. There are not as many formal project management procedures in place as the project would like. There is an Advisory Group but it has no responsibility; ultimately the Director is responsible to the Director of Research Libraries. The managerial quality assurance procedure is through the Advisory Group.

Neither pilot nor feasibility studies have been undertaken to assist project management. Some time and motion studies have been carried out in-house by the digital imaging unit (with a background in film processing). The project is starting to use MS project manager for scheduling. Who does what work was based on early arbitrary staff decisions.

Digitization is both carried out in-house and is outsourced. The more standard, bulk material is outsourced for efficiency, while the more specialized, rare and fragile material is digitized in-house. Some equipment was available in-house and other equipment was bought in. The decision on which digitization process to adopt is based on flow or batch processing wherever possible. Flatbed scanners (2xHP Deskscans) digital cameras (SLRs with digital backs) and high-end professional cameras (Kontron) are all used by the project. Guidelines for data capture procedures have been established (an external company for calibration and practice manual). The benchmarks used are Kodak gray scales and color charts.

24.3 Human Resources and Training

The project employs one director (100%), potentially four metadata specialists (two in post 100%) plus three part-time (1.5 FTE, 50%) twelve curators (10% max), eight digitizers (five in post 100%), technical support/development staff (three in NYPL information technology group) one FT 100% on project.

Almost all the staff have library backgrounds, the technicians are fully qualified, the digitization staff were trained on the job and the manager has a library photo background. About one third of staff were re-deployed from other areas. Both internal and external advice was available on the technical aspects of digitization (handling, light, heating).

The training needs of the project team have not been assessed but areas where training needs have been identified are project management, preparation and handling of materials for digitization and the technical

operation of digitization equipment. All members of the project team have or will be engaged in training and this has been organized in-house (own and external consultants), and via external courses, independent study and learning on the job. The training has so far met the needs of the project.

The project has identified that training must take place to give coherence to the whole project. Production staff are being taught to see the wider picture of the project they are working for, and the importance that metadata, and other aspects in order to give full and meaningful delivery of the objects.

24.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and owns the copyright to the original materials. The copyright or rights status of the final digital deliverable is declared in a copyright statement on the web. Although the Library digitizes material in copyright (under the legal provision for libraries and with the owners' agreement) the project does not do this.

Users of the digital deliverables are allowed to make printouts on paper and film, burn to CD, DVD etc., and download to a PC, LAN or WAN as long as it is for personal use. Users can download and view thumbnails and lower quality images, download and listen to less than 30 second sound samples and full length compressed sound, and download and view samples and lower quality digital video clips for moving images. There is much discussion in the project and Library at present about the use of electronic management systems, such as watermarking.

The project has a conservation procedure for the original materials whose condition is investigated by conservators. No conservation work has been undertaken yet. The project has not modified, degraded or compromised any material to carry out digitization but much was already dis-bound. Risk assessment of the material during preparation for digitization was undertaken by an external consultancy and made use of the Head of Conservation's guidelines (which were based on Library standards). Special equipment and processes specified include book cradles and exposure limits. Materials are prepared by curatorial or preservation staffs prior to digitization and are sometimes monitored by them during digitization, but in most cases this is avoided. No restrictions are placed on the originals post-digitization but this is a proposal for future material.

Cataloging systems in place prior to digitization include catalog and reference systems (CATNYP) at the collection level as well as bibliographies and handlists. Identification and metadata information is used from these records in the digitization process. The project does not have access to all the relevant information (e.g. cataloging) and has to locate some core reference or source material. The project does not intend to alter the originals for the digitization process. Material has been rejected for incoherent paging, incomplete works, and micro-images of inappropriate density and blurred pages. Where possible, the project has digitized from originals. Intermediaries that have been used are slides/35mm or 4x5 transparencies, photographic prints (the project might use microfilm), 78rpm discs, wax cylinders and video copies from film.

The project does not catalog the original material *per se*. A record is created for the digital surrogate, now based less on Dublin Core, since the intention is to create a bibliographic record with rules of inheritance: a system of such hierarchical resemblance is not premitted by Dublic Core. Tools for controlling data values are LC subject headings and perhaps name authority files in the future. The metadata details are attached to an open source, so as not to embed it in the actual object. This is a model borrowed from Harvard, in which XML information is included with the object, and will take an estimated two years to be put into general practice. The digitizer and an information professional create the records. This metadata record for the digital deliverables is then held in a separate catalog in electronic form on an intranet server. It will eventually be available on the internet. The records for the digital deliverables and the original digitized materials are independent of each other.

24.5 Format, Resolution and Compression of Digitized Materials

The project has so far had no requirements to adopt standards for handling textual materials. The TIFF file format is used for capturing and preserving images. Delivery is by TIFF, JPEG and SPIFF-file, from one compressed file and its sample, rather than from derivatives. The project is also considering JPEG 2000. MrSID is used for large images, while the Lunar image processing system (which can do what MrSID does in real time) is also being used for managing service copies.

Image, sound, moving image DPI, bit-depth, sampling rate and compression details are not known.

The quality control procedures in place for the digital deliverables are random checks that vary from material to material with a total check on problem material. Metadata quality control procedures involve a second person check and random samples on outsourced materials. These quality control procedures have had a knock-back effect on scheduling.

Users do not have to pay for the use of the digital deliverables, but have the opportunity to order high quality images off-line. Users will eventually have access to the digital deliverables through an open- access catalog plus the materials, but at the moment it is restricted to in-house use. The system offers searching on creator and title, but is limited by data format. Searching and browsing metadata are handled by metadata stored in an Informix database plus sequel calls and Java.

Apart from searching and browsing, users are able to manipulate images by zooming in and out using MrSID and Lunar.

Usage levels are not applicable at the moment but will be monitored by automatic data capture.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, conferences and meetings (the project is obliged to hold a research conference each year). The project does not know which medium has been the most effective.

24.6 Evaluation, Funding and Long-term Sustainability

No front-end, formative or summative evaluation has been done but this will be undertaken when the research post comes in. In the meantime the project did not want to hold up digitization.

The project has funding of $5 million over five years, the main sources of which were private donors to the NYPL. The project is monitored by an Annual Report and a mid-year verbal report. The funding body requested a project management plan and cost models, and is expecting workflow reports as part of the Annual Report.

New material (and its associated metadata), metadata changes, user interface updates and file format changes will be ongoing but it was not known at what frequency.

The project is about to start work with the Mellon Foundation on a preservation strategy that will include the archive server as one of its components along with migration of data.

The digital deliverables will be available indefinitely and the project's longer-term sustainability is reliant on library funds generally, rather than self-generating funds. Loss of the digital deliverable would be a matter of concern.

---

[7] An original interview had been carried out with the previous director Michael Alexander in September 2000, but it was felt that there had been sufficient change to warrant a new interview with Barbara Taranto. The original interview had focussed on the digitisation of the 60,000 images, but the emphasis is now on the creation and development of a Digital Libray Program.

valid xhtml 1.1
abp~04/02

25 National Library of Norway

HATII interviewed Svein Arne Brygfjeld, Adviser for IT and Innovation at the National Library of Norway's Office of the National Librarian, on October 27 2000. The Office of the National Librarian is a co-ordinating unit for all activities within the National Library. Among the programs run by the Office, is the National Broadcasting Corporation (NRK) Radio Archive Project. The NRK was formerly the only national broadcaster, and the project intends to preserve the material it produced. In this way, the resources that the project is able to archive and make accessible will be of significant educational, cultural and historical value.

25.1 Organizational Digitization Program and Policy

The National Library of Norway is divided into three main structural components:

- the office of the Librarian in Oslo (c. 10 employees);
- the existing analog National Library operation in Oslo (c. 130 employees);
- the National Library unit in Mo i Rana (c. 200 employees, established in 1989).

The question of a collection survey assumes that a formal digitization strategy exists, which is not the case. Rather, the National Library is in the process of learning from experience towards establishing a strategy. It now has several years of experience gained at its unit in Mo i Rana, to develop in the digital domain. It was in early recognition of the strategic importance of digital activities in large-scale multimedia information management that the substantial, specialist unit in Mo i Rana was established in 1989. As a testament to the forward-looking attitude of the National Library, the main argument for establishing the new unit was the relative maturity of IT at that time which has, of course, now been superseded. The unit at Mo i Rana consists of state-of-the-art IT activities, a Media Laboratory (c. 50 people working on restoration, preservation and digitization activities) and a sound and image archive (c. 20 information professionals).

Another issue which arises in connection with overall policy, particularly when considering audio and moving image material, is what is meant by digitization: do we include, for example, the copying of fragile, nitrate-based filmstock to digibeta format after restoration as a digitization activity? Overall, as far as collection surveys are concerned there is a mixed picture. There are several projects and activities across all media types being undertaken at Mo i Rana. For some projects surveys have been carried out, but the decision to do so is made on a project-by-project basis. For those projects that have had collection surveys, relevant people would be made aware of the survey and encouraged to participate. These decisions are made in a pragmatic way: those staff members who are available and have relevant expertise will be consulted.

In terms of passing on advice for good practice in this area, it is important that the needs of partner organizations are taken properly into account. The National Library carries out several partnership projects with other organizations, e.g. it is currently involved in a project for networked-based learning in primary schools. The partner organizations should also be encouraged to contribute to collection surveys.

The National Library's view of its digitization operations is of general "activities within the digital domain", rather than the narrower concept of just digitizing its own holdings. Many of its activities are carried out in co-operation with others in the community.

There is no formalized strategic policy statement in this area, rather the National Library's priorities are, to some extent, derived from its funding remit. The National Library is funded by the government to serve the interests of "research and documentation" — this remit can be and is interpreted widely by the Mo i Rana unit. Activities within the digital domain tend to facilitate enhanced access to materials and digitization, and the dual aims of enhanced access and social inclusion in practice drive policies and thinking about digitization.

The organization is not monolithic — rather the impression is of a flexible and responsive *modus operandi*. The Mo i Rana unit investigates what technology and methods are available for different projects and what is possible in its different areas of operation. Many of the activities are research-led in order to gain experience of

whether they are going to work on a large scale. Even for pilot, research-led projects, the tendency is always to digitize a fairly large chunk of the collection. There is considerable freedom, within the interests of research, to work in different domains and to pursue various methods and technologies.

The largest obstacles both to planning and building digital deliverables are perhaps organizational aspects within the larger institution, including the culture of the traditional analog library world. Although organizational aspects are most important here, technology is also relevant. Technology is always a little behind what is needed, especially for larger scale projects.

All of the criteria listed (with the exception of IPR) are relevant in the selection and prioritization of materials for digitization. Generally speaking the highest priorities would be:

- teaching and learning potential (e.g. what the National Library can do with the materials — can they be made available to primary schools, for example?);
- research significance;
- enhanced access;
- preservation (this criterion becoming more and more important over recent years and higher on the list of priorities).

IPR is not understood as a priority in this respect. The National Library does not necessarily own copyright on materials, especially those deposited under the fairly recent (1989) legal deposit legislation, and so it is not necessarily allowed to make digital materials available.

Other organizations with which the Library works may have other priorities, e.g. the National Broadcasting Corporation (NRK: http://www.nrk.no) over the radio archive project.

Commercial exploitation is almost always via partners, e.g. the radio archive, where the reason for digitizing is that the programs will be re-used by either NRK or other radio stations. In another case of working with partners, there is a database of Norwegian music which the company concerned wishes to make available for commercial use on a commercial basis, but which is deposited with the Library.

Selection and prioritization criteria have changed over time, in that preservation has become a higher priority.

There is a great deal of co-operation and partnership with others, with the National Library in the role of the provider of expert technical services across the board. All type of organizations (except foundations and charities) and all types of regional and international co-operation are undertaken. The National Library is proud of its achievements in this area and its record of co-operation with partners.

The best advice for those embarking on partnerships is to ensure that there is the best possible understanding of the roles and interests of each collaborator and that all involved are aware of these roles. There has to be a common understanding of the focus of the project. Personal relationships are also important — there must be attention to trust, respect and understanding of the professional skills and organizational aims of others. It is vital to try out working with partners and establish ways of working together via small-scale pilots.

The overall program is ongoing. It started with the setting up of the Mo i Rana unit in 1989 and is continuing indefinitely.

All of the items listed apply with teaching and learning, public access, preservation and research (or keeping a technological lead) as high priorities. The National Library is proud of its research-based record in choosing appropriate technology and looks upon this area as a successful one.

Revenue generation may be on the agenda of partner organizations, rather than the National Library's. Response to previous demand for digital materials has not yet come from users, but there is a demand from partner organizations.

There is an overall statement of intent, or mission statement, on the National Library's website. All of the individual points listed may apply to specific projects, but not necessarily to the overall program. The mission statement refers to the need to establish a framework for digital libraries and collections which make information available to everybody, wherever they are, in a democratic way. The emphasis, again, is on access and inclusion.

The thinking is to establish a long-term digital library with core functionality, and then provide different "windows" into this information.

At the moment, there are many different digital collections and media types, such as 100,000 still images, a large collection of newspapers, audio (including 400 radio programs), entire collections or radio programs on Norway in World War II from the US, “Spirit of the Vikings”, examples of poetry with poets reading their own works. Taking the example of the NRK radio archive, one “window” into the information would be for use by NRK staff for their professional activities; there might also be a public view; equally, there might also be a view for educational users from an authored, guided website where paths are more clearly defined. The way that the information is presented will vary a great deal but it is of greatest importance to get the core functionality right in order to facilitate different views of the materials.

A variety of source material is being or has been digitized with the exception of textiles and artifacts. Concentrating on audio-visual materials, all types listed are being digitized. The nature and format of the materials are similarly varied, and include TV and nitrate-based film stock. There are some entire collections, as well as some samples.

Digital deliverables are intended to be re-used and re-purposed, for example in the networked learning for schools project. The preference is for networked delivery rather than CD-ROM, but other delivery modes are not necessarily ruled out where they may be appropriate.

Tools for representing content are used as appropriate, e.g. for audio-visual materials, MPEG is used.

Dublin Core is used for describing content, while a variety of other standards are used from the original cataloging including non-standard systems, e.g. in broadcasting or film where institutions have often adopted their own house styles.

In relation to the use of standards in general, it is worth noting that rewriting metadata — even where it is desirable — is extremely time-consuming and will only be carried out when the collection is small. In most National Library projects it would not usually be done; rather existing cataloging is transferred to Dublin Core. For example, in the radio archive project the National Library had to rely on the previous cataloging metadata from NRK which was sometimes of poor quality, but resources were not available to improve it. In this case new metadata-creation would have to be done by actually listening to the material — ideal, but not realistic.

URN, as a persistent world-wide unique identifier scheme, is also used for identification in the NRK project. NRK has agreed to use Dublin Core. Agreement on standards makes it much easier to interoperate with “publishers” such as NRK.

Target audiences are primarily those in “research and documentation”, as in the National Library’s remit, but in practice they are very wide, including all the target groups mentioned. As noted above, the Library’s thinking is to provide different, tailored “windows” onto its repository of digital materials, and these windows would be constructed on a case-by-case basis, as appropriate to the particular needs of an audience.

25.2 Project Management and Planning

Project management is done in-house. There was little previous culture of project management at the National Library before the Mo i Rana unit was set up. Mo i Rana staff came more from different backgrounds, including commercial backgrounds, and brought a culture of project management with them. The Mo i Rana unit now sees organizational/project management issues as a successful area of its operations.

The organization of the project management responsible for formally constituting projects, can be summarized as follows:

- project leader (who is given and is enabled to take responsibility);
- steering group (small group of 4-5 which makes strategic decisions, directs the project and decides priorities, meets 4-6 times per year);
- reference group (including one member from each professional domain represented in the project (e.g. technical, or a metadata specialist), meets x1 per month);
- budgetary procedures and formal follow-up procedures should also be in place.

The main advice regarding project management is about inclusiveness and delegation of responsibility: it is important to ensure that all the relevant people are included and to place responsibility with the right people — project members should know who has responsibility.

Feasibility and pilot studies are carried out for individual projects. These tend to be informal, and based on

experience, rather than formal benchmarking/time and motion studies. Decisions on delegation are made by relevant team leaders on the pragmatic basis of which staff members are available (there is little staff turnover).

Digitization is carried out in-house, with the National Library's own installed equipment. The National Library acts for others as the expert provider of services and expertise.

25.3 Human Resources and Training

The profile of staff in Mo i Rana tends to be quite different from that of the National Library in Oslo: they tend to be younger, technical, and IT-orientated. Approximately 10% of IT staff have Masters degrees, and there are few conventional librarians. Technicians in the Media Laboratory are unlikely to have had previous technical training as no training to that degree of specialization is readily available; rather, training is done in-house. The technical department has a very high degree of skill — e.g. they have recently constructed from scratch, a player to play wax cylinders more efficiently.

Training needs are assessed project-by-project and training is often carried out on the job. External training and work experience are also used: people have been sent to UK and US international corporations and laboratories, e.g. Kodak in Rochester, particularly for those working with film and audio.

25.4 Project Life-Cycle Processes and Procedures

Legal deposit was only introduced for the National Library of Norway in 1989. The legislation is focused on the information itself, not the information carrier and legal deposit covers all media types including web space, TV and radio. The National Library does not always own copyright and copyright does not necessarily loom large in the decision to digitize, due to this legislation.

User access to digital materials is varied: for example, researchers have free access, but for materials for schools, or otherwise re-purposed materials some payment may be made.

No electronic system is currently used to control copying. Methods of managing rights in an object-based way are being researched to control usage on the networked-based learning project. This research is investigating a model which "wraps" the digital object in a coating of rights-management layers.

Mo i Rana is very well equipped with high quality film and sound laboratories. The curatorial staff make decisions about what is digitized, taking into account conservation needs, preparation of material, risks and any need for special equipment and handling (such as with nitrate-based films).

Access would not normally be restricted after preservation/conservation.

A variety of cataloging or reference systems is used for original materials prior to digitization from standards such as MARC to in-house systems. As noted above, even where it is desirable, re-writing metadata is not normally undertaken, rather existing cataloging is transferred to Dublin Core.

Digitization is used to preserve materials (e.g. fragile sound or moving image material) and these types of materials would not normally be rejected for digitization on grounds of poor original quality. Rather, curators would decide that fragile and vulnerable materials would be prioritized for digitization. All the forms of reproduction or intermediaries for audio and moving images would be used.

The degree of sophistication of metadata for digital objects depends on resources and quality of the original cataloging. The records for digital deliverables and original digitized materials are stored independently of each other. Metadata records are kept separate from the main National Library Catalog and are available on the internet.

25.5 Format, Resolution and Compression of Digitized Materials

In general, audio material is captured and preserved in WAV format for high quality. For access and delivery (depending on circumstances) versions in MPEG layer 3, MPEG 2 (for digital radio broadcasts) and RealAudio are used. The audio sampling rate for capture and preservation is 48kHz and can be 11, 44 or 48kHz for delivery. The bit-rate at all three of these stages can be 16, 20 or 24. The sampling rate for old originals, for

example from wax cylinders, is 16 bit stereo at 48 kHz. Where the originals are of better quality 24-bit stereo is used. The average file sizes are 0.5 GB for capturing and preservation and 25 MB for delivery. Bit-rate compression methods used are MPEG 1&2, layer3 and Real Audio for delivery only. The aim of compression is to reduce costs, improve access, enhance usability and decrease storage requirements. Post-capture editing is carried out.

Digitization from 35 mm film is done using a film scanner. The film scanner produces a digital video signal (270 mbit per second), which is processed to minimize noise and distortions. Then the film is stored as a high quality copy on digital beta and a “low quality” compressed copy made. The thinking is that digital formats are not yet good enough, and resources not big enough for huge storage requirements of disk copies. According to needs and circumstances, digital versions may be made in MPEG 2 at 6 mbit per second and MPEG1 at 1.5 mbit per second, but the digibeta copy is the main preservation medium.

Quality control procedures depend on context and the process and material in question, such as knowledge about the source of the original. Random checks are made on the quality of the end product compared to the quality of the original.

The catalog and certain materials are now available to users on open access, but this is not on a large scale and user statistics are not yet being collected. As noted above, search and browsing facilities will differ according to the user profile.

Availability of the radio archive is still under discussion. Some will not be made available, as it contains confidential material. Again, there is a mixed scenario. Some partners such as NRK may charge for access to their materials. Other partners, such as record companies selling digital copies of music on the net, will charge. Record companies have deposited audio material which can still earn money via CD distribution, but the National Library retains rights to make material available to core users under its remit of use for research and documentation.

Potential users are generally informed about the collections on the website, but also through special events for particular projects, for example a special launch of materials about Norwegian immigrants leaving for the US, which was also launched in the US.

25.6 Evaluation, Funding and Long-term Sustainability

No evaluation data is available for the overall National Library program, although individual projects may carry out evaluations.

The level of government funding is in the public domain but may be difficult to interpret generally. Investment in infrastructure is likely to be between NOK 7,000,000 and NOK 10,000,000. It is more difficult to calculate man-hours. Everything is more or less funded by the government, but there are situations where costs are shared, as in the NRK example where the digital radio costs are shared 50/50 by the library and NRK.

Preservation strategy for simple and single data types (i.e. discrete units not cross-referenced data) such as audio recordings or moving image recordings, is not to store on carriers like CD-ROM or DAT tapes. That is, not to store onto formatted information carriers, but rather to store on disk in the data repository, with migration as the key. The strongly held view is that migration is preferable to emulation, which has not proved successful and at the moment it is too difficult to maintain over time. Alternatively, it is possible to build up a technical museum, but this provides too many challenges when machines stop working and the knowledge of how to use and maintain them is lost.

There is no exit strategy — on the contrary, the digital activities at Mo i Rana are aiming towards building a sustainable digital library. Loss of the digital deliverables would be a matter of concern.

26 Survivors of the SHOAH Visual History Foundation

On January 24 2001, HATII interviewed Douglas Greenberg, CEO of the Survivors of the SHOAH Visual History Foundation (VHF); Sam Gustman, Executive Director of Technology; Karen Jungblut, Director of Cataloging; and Kim Simon, the Director of Archival Access and Community Relations. The aim of the VHF was to record and preserve the testimonies of Holocaust survivors, and to this end conducted thousands of interviews every year, until the collection of over 50,000 testimonies was complete. The Foundation has the substantial task of cataloging, storing and disseminating the video recorded material, in order to make it available as a crucial resource for education and research, while preserving it as a principal source of historical evidence. In addition to these functions, the very process of digitizing a collection as vast as that undertaken by the VHF prompted the development of new technologies in the field.

26.1 Organizational Digitization Program and Policy

The Survivors of the SHOAH Visual History Foundation is a unique organization in that it has one area of collection and one desired objective which together form its single *raison d'être*.

It is necessary to describe the impetus that led to the creation of the VHF to understand the whole process and details. In 1994, after filming Schindler's List, the film director Steven Spielberg was so frequently approached by survivors wanting to tell their stories of the Holocaust, that he was determined to record as many of these testimonies as possible. Thus the VHF was founded with the remit to interview, record on video and digitize 50,000 unedited testimonies from survivors from all over the world. This mammoth task has been undertaken with extreme commitment and vigor by the dedicated staff and in four years the 50,000 testimonies were recorded – indeed, at the time of this visit, the total was just under 52,000. To give an idea of the amount of material, it would take over 13 years to watch all the interviews.

There was no collection survey as such, merely the remit to interview and record the survivors' testimonies, to archive them and to make them accessible to diverse communities. The survivors come from all over the world, representing 57 countries and 32 different languages. All survivors are interviewed in their preferred language.

The survivors encompass the range of people persecuted in the period 1933-1945 and include:

- survivors of ghettos and camps;
- child survivors (including children in hiding, in camps, and in ghettos);
- those who survived by hiding/those who lived under false identities;
- refugees;
- members of the resistance (including partisans);
- rescuers and aid providers;
- Jehovah's Witnesses;
- Sinti/Roma (Gypsy) survivors;
- homosexual survivors;
- survivors of eugenics policies;
- liberators and liberation witnesses;
- war crime trials staff members and witnesses;
- other eyewitnesses.

The organizational policy is to take these recorded interviews and store them digitally as well as in analog form to allow preservation and retrieval of this unique collection.

From the beginning of the organization, the strategic thinking was to devise a system that would enable this policy to be put into practice. The system was conceived from the outset and this is one of many factors why the foundation has achieved its main goals so far. The system was designed to consider all the stages and phases in the collection, storage and retrieval of such material and so every step is taken within the larger framework.

Their target audience is wide and varied, covering all levels of education from K-12 to graduates. The driver is of course the unique material and the format in which it is presented. While there are archives of textual materials from survivors of the Holocaust, the VHF is unique in that the testimonies are spoken by the individual and thus heard and seen by the observer. This was the principal motivation behind the concept of the archive, to enable future generations to see and listen to the testimonies from those that survived the Holocaust.

It is not really possible to say whether this collection is a sample or the whole. It can never be the whole as such an undertaking would be overwhelming. However, it is such a large and expansive archive, with one subject only, that to call it a sample does not do it justice. As the Foundation has only one type of source and is unique in its format, it has not set out to be interoperable with other catalogs and archives. The archive is intended to be searched by users at particular centers around the world and is not intended, for instance, to be linked to a larger system of museums. However, in the creation of the catalog it has consulted the Library of Congress about thesaurus and name authority files. Due to the nature of the material, much of the catalog keyword, thesaurus and naming conventions had to be created to allow for the classification. The NISO Z39.19 Standard for Structure and Organization of Information Retrieval Thesauri has been used for the catalog and the retrieval of the information from the testimonies. They have not used standards for the processes because these did not exist for the work they wished to do.

The priorities for collection and digitization are many, with access to the materials and historical and cultural value being the main criteria. Education (teaching and learning) and research are also high priority, with co-operation with local educational organizations to use the materials in syllabus to educate the current generation about the events of the Holocaust. Other criteria include:

- inclusion in repositories and universities worldwide;
- historical and sociological research;
- development of educational materials, including documentaries, CD-ROMs, and publications;
- creation and enhancement of museum and other public exhibitions;
- international development of local and regional videotape libraries.

The primary source nature of the archives makes this a unique resource.

26.2 Project Management and Planning

The project management for VHF is completely in-house. As the project and organization were conceived as a whole, the structure has been in place since the beginning.

A board of directors helps guide the institution. The CEO oversees the various departments and all departmental heads direct their staff accordingly.

At the time of the visit, cataloging the interviews was the main task, with half of the staff and 30-40% of the budget devoted to this area.

Each department head manages his or her department in accordance with goals and timetables. A weekly track on cataloging reveals any areas that can be improved upon and identifies incentives for staff to catalog more quickly.

There was no pilot project as such, but careful planning and development of the system took place from the onset of the project. In 1994, the interview scheduling system was created. The digitization system was completed in 1996 and the catalog was developed in 1995-98. The networks for distribution have been developed throughout 2001. The organization works on industry project management procedures and has a top down approach.

All the digitization of the material is carried out in-house, as are all of the processes apart from the interviews, which are naturally conducted in the survivors' homes.

26.3 Human Resources and Training

The personnel in VHF are an extremely dedicated group of people. There are 150 staff covering areas including interviewing, cataloging, fund raising, information systems, shipping, translation and outreach. These staff come from a wide variety of backgrounds: the catalogers are librarians, archivists, or historians; the interviewers come

from a similarly wide range and include survivors themselves, historians and relatives of survivors. The staff employed in the day to day running of the institute, such as shipping and information systems, are all people skilled and trained in their appropriate areas. This is one of the strengths of the VHF; much of its success derives from the dedication of the highly committed staff, many of whom were identified as key players in their fields and asked to become part of VHF. Employing staff with particular skills in one area, who then bring their understanding and experience to the role in the process, rather than having to re-deploy and re-train existing staff, has been an immensely successful strategy.

Interviewers go through a rigorous training process to help them and the VHF assess if they are suited to the task. The content of the interviews makes the process emotionally difficult and not all are capable of dealing with this. The interviewers and video recordists are trained in interview techniques as well as being counseled to deal with the emotional aspect of the task. Prior to the interviews the interviewees fill in a Pre Interview Questionnaire (PIQ), which covers the details and facts of the testimony, such as names, places and dates. This PIQ has a twofold purpose: to assist the interviewer in more fully understanding the survivor's testimony and also as the basis for the retrieval and cataloging of the finished recordings.

The catalogers are trained in the use of the catalog, both practically and semantically.

Most of the training is done in-house by the appropriate team. Catalogers have the difficult task of listening to testimonies in detail day after day, and VHF have counselors and advisers available for the catalogers to speak to should they become overwhelmed by the details of the testimonies.

26.4 Project Life-Cycle Processes and Procedures

The copyright of the materials from the interviews belongs to the VHF. Each interviewee signs a release that gives the Foundation the rights to distribute the testimonies for educational purposes. However, copyright becomes an issue, since the Foundation has a duty to protect the testimonies from misuse and misrepresentation. To this end they have identified four categories of users:

Foundation authenticated – these are early requesters who are using the archive within the Foundation itself. They are known to the VHF which has approved their use of the materials for research.

Location authenticated, where the VHF knows where and who the users are through the interfaces at identified museums There are five initial repositories:

- Yad Vashem, The Holocaust Martyrs' and Heroes' Remembrance Authority, Jerusalem;
- United States Holocaust Memorial Museum, Washington, D.C.;
- Museum of Tolerance, Simon Wiesenthal Center, Los Angeles;
- Fortunoff Video Archive for Holocaust Testimonies, Yale University, New Haven;
- Museum of Jewish Heritage – A Living Memorial to the Holocaust, New York City.

These users would have pre-requested access for research in their discipline.

Location unauthenticated, when the VHF knows the location but not the identity of the user. This may be through an exhibit in a museum or restricted use of the catalog and finding aids.

Globally unauthenticated, potentially via the internet but as yet this is not a reality and much work will have to be done on the authenticated users to understand how the materials are used and how they should be accessed.

All users have to apply to use the archive and sign a non-disclosure document before being granted a license to publish.

The VHF will work on the authenticated users before looking at the unauthenticated possibilities to ensure that any procedure works. This will also enable the evaluation of the interface and finding aids and catalog. They have to balance the freedom of information against the integrity of content. They have a large legal department working on documents to ensure that the testimonies are protected from misuse.

26.4.1 The Catalog

The catalog for the archive is the key to accessing the materials. The remit of the Foundation was to record the interviews on video so that the survivors could be seen as well as heard. Unlike many oral histories, these resources are not transcribed for searching and retrieval.

The PIQ is the first step to accessing the materials, enabling users to search for names, dates and places at a metalevel which will result in focusing on a selection of the interviews. However, the final goal is to allow users to search for keywords and phrases, before actually seeing the clips from the interviews that relate to each search item. The Library of Congress Subject headings were too broad for this intention, however, where possible, national standards are used for this purpose and adapted where necessary.

A team of historians decided on the information for the short form PIQ. These have to be translated from the language they were completed in by the translation department so that, as the initial finding aid, they can be used by the English speaking community at first, although eventually the catalog and finding aids should be accessible in all 32 languages for users. The thesaurus used in the catalog is monolingual but has to be expanded to 32 languages.

This is done by dividing up the testimonies into discrete segments, identified by the cataloger. These can be anything from one to twelve minutes, although the average is around 3.52 minutes. Each testimony varies in length from one to seventeen hours, with the average at two and a half hours. The segments are marked on the digital version, so there is a start and end marker. These segments will be discrete items of information, such as arrival at a concentration camp or the description of a person. The cataloger then associates this segment with keywords that describe the segment. The list of keywords is content driven and the catalog grows as new keywords are discovered. The catalog department has teams of catalogers and researchers that review the testimonies and run quality assurance techniques on the cataloged testimonies. The researchers also oversee and investigate the new keywords that may have to be added to the catalog, where a testimony may bring to light an as yet uncataloged idea or description. Weekly editorial boards look at the methodology and the vocabulary. There are also group meetings which can help shape the methodology. The cataloger proposes an addition, which is researched and then authenticated. All updates are logged in a database to track changes. There are approximately 21, 000 keywords that are related to each other.

At the time of this visit 4,146 testimonies were fully cataloged and 42,682 short form PIQs were completed.

The segmentation marks on the digital video are associated through the metadata to the original object. The metadata scheme for the digitized video is large and complex and while various standards were considered, the VHF has created its own system to inform about the digital process, format, catalog link and administrative details. The metadata is populated automatically and the barcoding of the tapes underlies the various formats the video will take.

26.5 Format Resolution and Compression of Digitized Materials

The interviews arrive in Beta SP format and are then entered into the index that organizes the tapes. Each tape is barcoded to associate it with the PIQ and thus the interviewee and interviewer. The tape is then copied to digital Betacam for preservation, to VHS for in-house use, to VHS (or other appropriate video format) for the survivor's copy, and to 3MB per second MPEG1 copy. Each hour of video takes approximately 1.5GB and the average interview is 2.5 hours long. Compression using MPEG enables quicker access and reduces storage.

This process takes place automatically in the video transfer room, where 16 machines can copy simultaneously to four different video formats as well as the digital version.

In some interviews, survivors show objects to the camera. These are extracted from the digital video and stored as JPEG images attached to the MPEG version. No text is associated, as the interviews are not transcribed.

Users will see the MPEG format in the user interface.

Preservation copies are made and stored in separate temperature controlled vaults in California and New York. This double preservation ensures the longevity of the tapes.

26.5.1 Retrieval

The system for storage and retrieval is simple in concept but complex in structure. The user accesses a local cache at the repository for instant retrieval. Each cache is 1 Terabyte. If the information is not there, then the disc server (180 Terabytes) at the VHF is accessed; if the information is held on the actual server, then retrieval is almost instantaneous. The final step is for the information to be found in the tape storage and uploaded to the server and local cache. This is done automatically by a robotic arm that locates, identifies and loads the desired information, taking five to ten minutes. It is hoped that the local caches will hold the most commonly used data

and thus the user will not have to wait.

26.6 Evaluation, Funding and Long-term Sustainability

The user interface is at the beginning of its development and so evaluation of end users and interface functionality is just starting. There is an online survey for the interface at the five repositories and an analog one in South America and Europe to discover who the users are, how they want to access the materials and for what purpose. The VHF is aware that it cannot meet the demands of every potential user and that it has to protect the integrity of the material at all times. It is exploring how to meet its mission and at the same time serve public demand, which has raised large ethical considerations as well as issues of academic freedom. These have yet to be fully resolved. As the interface develops and the testimonies are made available at the repositories, this research will continue to identify and develop the authentication required for user access at various levels.

The Foundation has designed products with the intention to raise awareness and to help dissemination as well as creating educational tools. It has produced three documentaries and one interactive CD-ROM based on the archive, in response to users who wish to start accessing the material. The CD is targeted at educational organizations, in particular K-12. It contains four testimonies from survivors intermeshed with historical facts, timelines and maps.

The Foundation is a 501(c)(3) public charity. The initial funding of $45 million covered the production and collection of the interviews. The Foundation is now on the second phase and requires a further $50 million to continue collection and fully catalog the archive. At the moment, the Foundation has an annual running cost of $13.5 million. All the funding is from charitable donations.

The second phase is to ensure the longevity, retrieval and dissemination of the materials. This future production has been planned since the start of the Foundation and covers the areas of security and authentication. They are researching WANS (OC3 for 150 MB per sec retrieval). Internet 2 is a possibility, as is ATD Net. This is leading towards a global infrastructure of copies of the testimonies with the entire archive online. They will also continue to produce CD-ROMS, documentaries and other mass media products. As well as the global infrastructure, they will examine the use of one-off productions for standalone machines, perhaps geared at a specific area of research.

26.7 Comment and Conclusion

The Survivors of the Shoah Visual History Foundation is indeed an exemplary project that has devised and operated a simple but efficient system to collect, store, archive and catalog 52,000 testimonies from all over the world. It has a major remit and as such, has also had a massive budget to enable these systems and organizations to be realized. Much of its practice cannot be applied to smaller museums or archive projects, as the level of funding required will not be available to produce such sophisticated systems and technology. Nevertheless, there are basic principles that can be applied to any digitization project:

- employ staff who are skilled and talented in their area of employment. For example, if a database or catalog is to be created for metadata storage, find a skilled technologist who understands the systems, rather than, for example, a librarian with little specialized interest, which will result in the creation of a flawed final product.
- plan the system before any digitization takes place, envision the entire process and life-cycle and plan for that up to and including preservation of the materials;
- understand the collection, whether it is one type as in the VHF or a variety of materials in an archive or museum;
- hold the copyright for the material;
- have dedicated and energetic staff – the people at VHF were undoubtedly some of the most dedicated staff on a single project;
- use standards where possible but be aware that adoption and adaptation of purpose built systems will probably be inevitable.

valid xhtml 1.1

27 Thesaurus Musicarum Latinarum (TML)

HATII interviewed Thomas Mathiesen, the David H. Jacobs Distinguished Professor of Music and Director of the Center for the History of Music Theory and Literature at Indiana University, on January 10 2001. The Center initiated the Thesaurus Musicarum Latinarum project, which is an evolving database with the aim of collecting all Latin music theory dating from the Middle Ages to the Renaissance. The intended audience for the resources made available by the project, is principally those within the University community. In accessing the TML, users will be able to locate and retrieve the source text for educational or research purposes, while the original artifacts are preserved.

27.1 Organizational Digitization Program and Policy

The Thesaurus Musicarum Latinarum (TML) is a full-text database of Latin music theory, extending from Censorinus's *De die natali* through treatises of the seventeenth century. The database contains more than 5,000,000 words in 741 separate texts accompanied by more than 4,000 graphics. The TML originated in an idea in the spring of 1989 amongst twelve scholars active in textual criticism, codicology, early music, cataloging manuscripts and the history of music theory. These scholars had used the Thesaurus Linguae Graecae (TLG) and it was the TLG that provided the model for the development of the TML. A discussion amongst a larger group of scholars took place at the annual meeting of the American Musicological Society in Austin, Texas in October 1989. This meeting made a general commitment to the project and preliminary editorial and technical committees were established. Indiana University (IU) provided substantial funding to establish the principal TML center. A planning meeting held at Princeton University in 1990 established the basic parameters of the project, and the first texts became available later that year. The texts, which were not originally intended to be viewed online, were first distributed by request on LISTSERV, then by File Transfer Protocol (FTP). With the advent of Gopher and eventually the World Wide Web, the texts and graphics were restructured for online viewing. In 1992, the project received a $230,000 NEH grant in support of work on the creation of texts at the IU TML center and the TML sub-centers that had been established at the University of Colorado in Boulder, Ohio State University, Louisiana State University, the University of Nebraska in Lincoln, and Princeton University (an additional sub-center was eventually established at the Moscow Conservatory). A second NEH grant from 1994 to 1996 enabled the project to reach its goal of including 97-98% of all printed texts, to create the accompanying graphic files, and to begin work on manuscript sources. When the NEH grant ended in 1996 the database was receiving more than 1,000 connections per month. In 1998 the Center for the History of Music Theory and Literature (CHMTL) was established as an umbrella for various music projects developed at Indiana University, including the TML, and is currently in its third year of start-up funding.

There have been relatively few major obstacles to planning the development of a strategic policy for the TML. It was advantageous that the principles behind the project were established before the first text was created. The first obstacle to be cleared was obtaining copyright permission. Fundraising has also proved time consuming but has been successful. One main obstacle to building the database was the constant danger of being shut down as system administrators decided to terminate servers, change operating systems, and otherwise alter services on which the TML relied, quite often without any advance consultation. Therefore, a huge amount of time has been invested in moving data from a VM mainframe to a desktop machine emulating VM, an NT server (for the LISTSERV component of the project), a UNIX server, and finally a Linux server, and in resolving the associated political issues.

The principal criterion that guided the TML in its selection of material for digitization was that the most widely used printed texts would be digitized first. Therefore, provided intellectual property rights were secured, material that had the highest research significance had the highest priority. A feature of this selection criterion was that every edition of a text would be included, as distinguished from the TLG, which includes only a "best text" for each edition. A second selection criterion has been to digitize related manuscript sources to facilitate the production of new editions. These selection criteria have not changed over time, with the exception that the priorities have evolved. There is increased interest in extending the chronological limits backward to the second century and forward to the late seventeenth or early eighteenth century, as well as including material that is not

exclusively devoted to music.

The TML has co-operated with other academic institutions on an international basis, although the project was originally envisaged as a US undertaking. The TML was fortunate in this co-operation in that musicology is a small field of scholars well known to each other and able to collaborate effectively. The project director was also well established and financing was in place. The TML would therefore recommend to other institutions, that an essential factor for success is an administrative board that collaborates effectively.

The TML has no anticipated end date, as the number of potential manuscripts to digitize is huge and raises very different encoding issues as compared to printed texts. Although the pace of work has slowed, more than 100 new texts have been added in the last two years.

The primary purpose for the TML in creating the digital deliverables is research. The TML has produced a general description of the project that includes the TML's format, rationale, scope, significance and level of faithfulness to the original. The level of faithfulness to the original is embodied in the TML's "Principles of Orthography", published in the catalog of files called the *TML Canon of Data Files* (Lincoln: University of Nebraska Press, 1999).

The type of source material digitized includes:

- printed books, serials and similar bound material (published);
- printed illustrations;
- handwritten documents;
- intermediaries.

The printed material is all on paper and the manuscript material on paper or parchment. The digitized deliverables are intended to represent the entire body of material. It is not the intention of the project to re-purpose the digital deliverables.

Low character ASCII is used for representing content. The 127 low character ASCII set was chosen to achieve the maximum level of interoperability; because the TML was intended to be as editorially unobtrusive as possible, there was no need for any complex symbols or typography. For describing content, the text files contain simply a bibliographic header, database reference and the file creators. Page or folio breaks are indicated and non-Roman characters, in this case musical notes, are encoded: for example, a semibreve is S. The TML's encoding system, which is published in the *TML Canon of Data Files*, has proved applicable to all texts.

In the HTML versions of the texts, there is inevitably more markup, such as paragraph breaks <P>, and the use of graphics flags anchored to the graphic files; the graphics themselves are not embedded to ensure that the text files can be viewed in non-graphic browsers.

The TML did not look at guidelines for digitizing particular document types because many did not exist when the project started. The Unicode system was embryonic and the TEI just beginning. The only model available was that of the TLG.

The intended audiences for the digital deliverables are academics, scholars and graduate school students. Other user groups could use the texts but there is unlikely to be any significant demand from the general public or K-12 because of the specialized nature of the material. However, within the academic community the profile of users has been broader than expected, with medievalists, classicists and antiquarians making use of the material.

There are no limitations on the use of the deliverables because universal rights were cleared. In this process the project insisted that they were not replacing the original but were strictly concerned with the critical text.

27.2 Project Management and Planning

Internal advice was available on managing the digitization program but external advice was sought on serving the material (specifically in connection with searching capabilities and pre- and post-processing of searches). Indiana University's School of Music, as a professional school, has not shown much interest in the CHMTL, however, since 1998 it has contributed approximately one-third of the CHMTL's annual budget. The remaining funding comes from the Office of Research and the University Graduate School and from the endowed chair held by the CHMTL's director. The formal project management procedures in place are an external project committee and editorial advisory committee.

The TML does not carry out any feasibility and pilot studies but holds regular planning sessions. The TML has made changes to its delivery system following testing.

The creation of the graphic files occurred entirely in-house and the creation of the texts was divided amongst the various centers established by the project; with some text being sent in from outside the project. All the project's equipment has been purchased; it includes Apple and UMAX flatbed scanners, 21" monitors, CD burners, and DAT tape backup systems running on five CPUs with fast clock speeds.

An elaborate system is in place for creating the graphic files. The image is first captured in TIFF format, then edited and converted to PICT format and then to GIF. The GIF image is stored in UU encoded form on the NT server (for retrieval through LISTSERV and FTP) and in GIF form on the Linux server (for retrieval through the WWW). The preliminary TIFF and PICT files are not saved.

27.3 Human Resources and Training

The CHMTL employs a full-time director, who works a minimum of 20 hours per week on the project, although this level fluctuates. There is now also a full-time Associate Director of the CHMTL, whereas prior to 1998, there had been a half-time project assistant for the TML alone. During the period of the NEH grant, the project employed 2-3 student digitizers; the CHMTL now employs 3-4 students at 10 hours per week, although these hours are distributed amongst the TML and the CHMTL's other projects. The project staffs have a mixture of music graduate and medieval studies backgrounds. Ability in Latin is an obvious requirement, as is meticulous attention to detail.

Technical advice on digitization was available in-house. Training needs were informally identified in the areas of:

- technical operation of digitizing equipment;
- post-digitization processes.

New project assistants, associate directors and student digitizers have all received training from the project staff and through learning on the job. The training has met the project's requirements and the TML has been able to attract excellent student employees because it has paid quite generously.

27.4 Project Life-Cycle Processes and Procedures

The original holder retains copyright in the original materials. This copyright was cleared with the owner's agreement and in only one instance was a small fee paid. The TML claims copyright on the text compilations, table of codes etc., and defends the right of the TML to determine the method of data delivery. Users are allowed to make printouts of the digital deliverables on paper or film and download to a PC, LAN or WAN. No electronic management systems are in place to control copying.

For textual material users can view and download ASCII text files.

For digital image material users can view and download GIF files.

The TML does not have a conservation procedure for the original materials as this does not fall within the scope of the project.

Information for creating the *TML Canon of Data Files* (the catalog of the digital versions) is not derived from a catalog for the originals but the original material itself. As such, the record for the digital deliverable and that for the original object are independent of each other. Metadata recorded includes: the author and the title of the treatise (as given in the source); the incipit; the source and type (i.e., print or manuscript) of the data file; the filename, filetype, filelist, and size of the file in kilobytes; annotations; and the names of the persons responsible for entering, checking and approving the data. The digitizer and/or project director creates this metadata from the originals and with the exception of the incipit, the data is created to be Z39.50 compliant. The TML database is referenced in the University's library catalog but not at the text level. The *TML Canon of Data Files* is available on the internet and in published form (as noted above).

The project has not rejected any material for digitization. Photocopy intermediaries have been used for digitization, although the material did not exist only in this form.

27.5 Format, Resolution and Compression of Digitized Materials

The formats for retroconverted text-based digital deliverables are:

- HTML;
- txt (ASCII low level set plus notational markup);
- database.

The texts contained both Latin and non-Latin scripts. OCR was used as a conversion method for printed textual materials. The OCR software used was TypeReader and latterly also OmniPage. With clean text, the project achieved 98-99% accuracy, although the accuracy for texts derived from early printed books was well below this level. From this experience, the TML would recommend leaving the OCR decisions to the digitizer; nevertheless, even with poor OCR there are the advantages of avoiding such typical scribal mistakes as skipping lines or jumping from one word to a similar one elsewhere on a page. Keying-in conversion has been employed for manuscript material. Text file sizes range from 6 to 300KB.

For image material, the TIFF file format is used for capture and GIF for delivery. The capture resolution is 300dpi and the delivery resolution 72dpi; sometimes the images are enlarged to 150% for legibility. Capture is at 8 bits and delivery at 1 bit for line art (bi-tonal). The TML recommends being extremely careful with alignment and with threshold settings to achieve maximum clarity in the final image. The delivery compression used is GIF format to improve access and enhance usability. The average image file size is less than 64KB. The project does not retain the original scans in uncompressed form. The project carries out post processing on images using Ofoto (bundled with Apple scanners) and PhotoShop for clean up, de-skew and labeling. Ofoto still carries out some of the operations better than PhotoShop.

For others starting work in digital imaging, the TML recommends careful consideration of user needs and facilities, especially in terms of available equipment and bandwidth.

The quality control procedures in place for the digital deliverables are at least three sets of total checks on each text. The individual entering the data is expected to check and correct it before printing out a text to be passed along to the person responsible for proofreading and checking. This second person identifies and marks errors on the printed texts, after which each mark is counter-marked as the correction is entered into the electronic text. Then, both the marked printed text and the electronic text are passed to a third person for review prior to approval and addition to the database. Where there is a high error rate at any stage in the process, the text is printed once again and subjected to a second proofreading, as outlined just above. The final check and approval by the project director facilitates consistency in the quality control process. Lessons from this experience are that it has been difficult to get people to proofread character-by-character (rather than word-by-word) and to refrain from global search-and-replace editing. In general, the TML has discovered that no more than four double-spaced pages of 12-point text can be proofread per hour with an acceptable rate of accuracy.

Users have open access to the catalog plus materials. If it resides on an individual's personal computer, the TML has mainly been searched using the *GOfer* search engine, which the project licensed to distribute at cost. *GOfer* enables brute searches of as many as eight separate words or strings (controlled by various Boolean operators) throughout any defined set of texts, including such advanced features as soundex and proximity definitions. However, because the files are ASCII format, they can be used with virtually any search engine. For internal use, the project has developed *Eureka!*, a Windows application that has the same functionality as *GOfer* but currently without the soundex facility. *Eureka!* caches its searches, and it is therefore faster than *GOfer*, can do multiple sub-searches and retrieve previous searches. For searches on the TML website, CNIDR's *Isearch* has been implemented with a three-field form controlled by standard Boolean operators; because *Isearch* operates on indexed text, proximity searches are not currently available on the website. Users can, however, select a century context (or the entire database) and set the number of results to be displayed per screen. Versions of TML distributed on CD emulate the website (including CANTUS, a database of Gregorian chant manuscripts), and the CD can be searched using *GOfer* or *Eureka!* The TML averages more than 1,600 portal entries per month and processes on average more than 600 searches per month (search statistics are based on the number of times the script is invoked – i.e., each time a search is actually processed – not on the number of times the search form is retrieved).

Access to the entire database is free of charge. Users pay only for optional programs, such as *GOfer* ($20), or special media, such as the TML/CANTUS CD-ROM ($65). In theory, these are intended for single users, but this is not monitored or enforced. The charges were calculated to recoup the direct costs of the medium and distribution. Potential users of the database and its optional programs and alternative media are informed about their availability through a project brochure, website announcements, articles in print media, conferences,

meetings and an email distribution list (TML-L). Live demonstrations of the TML have proved effective, as was the original trifold brochure; email distributions have been least effective. Use of the TML increased exponentially when it was made available on the web.

27.6 Evaluation, Funding and Long-term Sustainability

The TML has not carried out any evaluations of the project (with the exception of the semi-annual and final reports required by the NEH) and has relied on informal feedback, of which they receive a considerable amount. If the project received complaints, it would carry out a formal review.

The TML itself has received $330,000 in NEH funding, and as part of the CHMTL, it has also benefited from the CHMTL's $90,000 start-up funds; networking and server costs have been absorbed by Indiana University. It is estimated that total project funding has been in the region of $500,000. The TML considers that it has received sufficient funding but the source of continuing funding remains unclear. To maintain the CHMTL (which includes the TML) in perpetuity would require an endowment of $3,500,000 (returning $175,000 per annum), which would provide sufficient operating funds and funding to bring fellows to the CHMTL to make use of its materials, engage in conferences, and offer instruction to graduate students. The TML's view is that the use of the standards it developed more than ten years ago has enabled it to grow and migrate data with very little need for revision or correction of its data.

New materials are digitized and added at the rate of 50 files per annum and this may increase. The user interface has changed very little over the years, and this fact is one of the features that users like. The TML does not have a formal strategy for preservation to ensure long-term access. At present the TML uses tape backup and stores CDs on and off site in controlled conditions for archiving purposes. In the early days of the project, backups were sent to Colorado; now, all board members have a copy of the data. The project is proud that in over ten years no data has ever been lost. In addition, all paper documentation is kept, including copies of the original sources and printouts of the electronic texts, on which are recorded all the alterations and checks. A longer-term preservation strategy is likely to rely on the (forced) migration of data. It is intended to keep the digital deliverables available indefinitely.

For long-term sustainability, the project is dependent on self-generating funds, but the TML has not secured the resource for this. Should resources prove insufficient, the exit strategy would be to break up the CHMTL's projects and move them elsewhere. Another possibility would be to provide the TML on a private server. The TML suggests that projects need to recognize that even institutional support is tenuous, that digital data is easily destroyed, and that some projects will not survive.

27.7 Conclusion

The TML is an excellent example of an elegantly simple project. It demonstrates the value of establishing clearly defined orthographical principles prior to digitization and formatting the deliverables to a recognized standard. In this case the standard may be the simple low ASCII character set, but it has proved to be both flexible enough to encode a range of material and migrate over different hardware and software systems for over ten years. Furthermore, the TML's encoded texts have been able to be accessed by very sophisticated search and retrieval software which proves that simple encoding is no barrier to greatly enhanced functionality. The TML did benefit from not having to digitize a wide variety of material or serve a diverse group of users. Nevertheless, one can extrapolate from their experience the benefit of adopting emergent text standards today, such as TEI and/or XML, to suggest that one does not necessarily have to implement the full complexity of these to provide long-term access to digital material.

valid xhtml 1.1
abp~04/02

28 Tufts University

Gregory Crane, Director of the Perseus Project at the History Department of Tufts University, was interviewed by HATII on September 21 2000. The Perseus Project is a developing digital library, which aims to make available a wide range of resources in the field of the Classics. The project is a reflection of Crane's interest in the relationship between the humanities and developments in digital technology. In digitally formatting classic texts, illustrations and artifacts, the project intends to increase the potential of the material for education and research, while also enabling broader public access.

28.1 Organizational Digitization Program and Policy

Digitization began thirteen years ago with a long term projection in mind. At first the project was unable to distribute the material and it was therefore collected on videodisc and 35mm film; now digitization work is done from slides. SGML tagging began ten years ago and later came into its own, as a very robust structure format.

The project did not conduct a collection survey but knew the corpus of classics texts; the sites were obvious and the key to museum objects was their accessibility or locating substitutes. The institution as a whole was made aware of this process and curators were consulted. The lessons learned from this were to get as many staff involved as possible and not to be too concerned about the details and documentation. If the project were to do this again it would not worry about people misunderstanding the nature of the project.

The process of identifying material was used to establish priorities for digitizing holdings. A critical mass of heterogeneous material was built with most early interest focusing on images as there were already Greek texts digitized elsewhere. These priorities were established by an editor in chief in conjunction with a subject specific editor (for example, in archaeology). These priorities were not formalized in a strategic policy statement but were internalized amongst the people on the project.

The objective this policy sought to achieve was the creation of a digital library to change the questions people asked and to reach a range of users and uses. The project feels it has been successful in these aims even though people projected expectations onto the project, which they would not have had for print material. Obstacles to achieving these objectives were the lack of guidelines, technology limitations and the lack of an inexpensive and accurate GPS system.

The greatest obstacle to planning and building the digital deliverables has been designing pieces that interact so that the whole is greater than the sum of the parts.

In the prioritization of materials for digitization intellectual property rights have acted as a filter but the main priority has been the material's teaching and learning potential, its research significance, potential to reach disadvantaged groups and social inclusion. Secondary priorities have been research into digitization strategies (considered synonymous with the project) and infrastructure cost reductions (as a means to an end). The selection and prioritization criteria have not changed over time.

The project has co-operated with archives, libraries, museums, academic institutions, corporations, foundations, charities, and government agencies. It has done so at institutional, local, regional, national and international levels. One lesson from this is to design projects with collaboration in mind and determine who will work with whom (e.g. LC rather than the Scottish Cultural Resources Access Network model).

The current status of the project is ongoing; it started in 1987 and has no anticipated end date.

The foremost purposes in creating the digital deliverables are teaching and learning, wider access and public access. Secondary purposes are research and experiment. The project has made an explicit statement of intent (through flyers, etc.) about its rationale.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- unbound printed documents;
- typescript documents;
- printed illustrations;
- artifacts;
- original artworks;
- intermediaries;
- hand copies;
- photographs;
- video recordings.

For images, the nature and format of the materials include 35mm slides and for texts, published books. The digital deliverables represent neither a sample nor an entire body of material; the digitized books represent two-thirds of surviving texts and the images are a representative sample. The project has re-purposed the digital deliverables in two versions of a CD-ROM, which has worked well. One obvious change the project would make if it did this again would be to make it possible to link to every object.

The following standards, guidelines or tools are used for representing content:

- TEI;
- CIMI;
- Unicode;
- JPEG;
- SGML;
- XML.

The project has not adopted any standards, guidelines or tools for describing content or controlling data values.

The program looked at other existing guidelines for digitizing particular document types but there was not much available at the start. It looked at beta code transcription for Greek and observed cataloging practice. The project rejected the beta code of the Thesaurus Linguae Graecae for page formatting and did not use any for control and display of page layout, such as HTML.

The following standards, guidelines or tools are used for representing structure:

- TEI;
- CIMI;
- SGML (for archive);
- XML (on the fly).

In relation to standards in general and suggestions for navigating between the ideal and the realistic, the recommendation is that XML, SGML and TEI are the least problematic. Also, only marginal benefit is derived from developing your own DTD; it is advisable to adhere to standards, get close to generic best practices and avoid excessive formatting.

The primary intended audiences for the digital deliverables are four-year college, graduate school, lifelong learning, distance learning and subject specialists such as classics, ancient history, or history of science. Secondary audiences include K-12 and community colleges.

Gary Marchiomini has undertaken an evaluation of the target audience. See:
http://ils.unc.edu/~march/perseus/lib-trends-final.pdf

The deliverables can be used by those outwith the primary intended target audience. The project has acknowledged the needs of those with disabilities via the W3C's "Guidelines for Web Site Accessibility". The project is anxious to minimize any limitations to use, but not all museum images are on the web due to the rights restrictions of museums.

28.2 Project Management and Planning

There has been external project management advice through a Technical and Academic Advisory Board (although not recently active). As a funded project Perseus can operate separately from the structure of Tufts.

The project has helped inform how Tufts thinks about IT, e.g. the development of its digital library, which will be the eventual home of the data. Formal project management procedures are a Steering Group of scholars from several institutions. The biggest project management challenge has been determining what people are good at and how they can contribute. Quality assurance procedures include spell checks, metadata captioning and review of front ends, which have all been successful.

A pilot study was carried out as part of a $500,000 contract to experiment with the creation of a digital library. This study covered scheduling, training needs, technical feasibility, user needs, workflow analysis, workflow piloting and technology forecasting. As a result of this study a scheduled collaboration between Harvard and Boston was shelved because of organizational over-centralization. No benchmarking or time and motion studies have been undertaken. The Editor in Chief delegated the work with the aim of optimizing the skills within the project team (for whom there are job descriptions).

Text digitization was outsourced for cost reasons, but is now moving in-house with the availability of high end OCR. The equipment for the project has been bought in. The adoption of the OCR digitization process was aided by looking at what the University of Michigan had chosen. Fujitsu flatbed scanners and Canon film scanners are used by the project. Guidelines have been established for calibration, resolution, etc. and vary according to the media. Color chart benchmarks are also used.

28.3 Human Resources and Training

The project employs one director (50%), one metadata specialist/curator (100%), one digitizer (100%), one photographer (for 12 years 100%), one or two technical support (100%) staff, three technical development staff (100%) and one 0.5 evaluation specialist (100%). Virtually all of the project staff have a classics background. None of the staff was redeployed from other areas. Advice on the technical aspects of digitization is available in-house.

Training needs were identified in project management, application of technical standards, preparation and handling of materials for digitization, technical operation of digitization equipment, post-digitization processes, metadata creation and digital preservation. All members of the project team have been engaged in training and this has been organized in-house (using own consultants), through independent study and learning on the job. The training has met the needs of the project.

28.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and owns some of the copyright on the original materials. The copyright or rights status of the final digital deliverable is declared. All methods of organizing digitization of material in copyright have been used (legal provision for libraries, owners' agreement, payment of a fee, under license and without formalities).

Users of the digital deliverables are allowed to make "fair use" copying of the digital deliverables. Users can view XML (to HTML) text, download thumbnails and lower quality images, and only view highest quality images. No electronic management systems, such as watermarking, are in use.

The project does not have a conservation or preservation procedure for the original materials. Investigations into the condition of the original materials is undertaken by curators and in some cases originals may, for example, be cleaned by them as part of the project. Some material is modified, degraded or compromised to carry out digitization. Risk assessment has relied on the work and advice of curators. Some special equipment is used to minimize stress on objects. Some materials are prepared by curatorial or preservation staffs prior to digitization and are sometimes monitored by them during digitization, especially in museums. No restrictions are placed on the originals post digitization.

Cataloging or reference systems in place prior to digitization range from published records to no catalog. Appropriate information is used from these sources during the digitization process. Both project staff and curators locate core reference material. Some early modern books have been rejected because it was not efficient to OCR or transcribe the handwriting. Reproductions or intermediaries (slides, 35mm, 4x5 transparencies, photographic prints, microfilm) have been used where necessary, but the material did not exist solely in reproduced form.

The project does not catalog the original material.

The digital deliverables are cataloged in an in-house relational database (Filemaker to Postgress to RDF). Dublin Core and RDF are the guidelines used in cataloging and the project is considering the use of MARC and USMARC.

Tools for controlling data values are thesaurus (e.g. of geographic names), museum catalog numbers, Thesaurus Linguae Greacae numbers, subject languages, Getty Art and Architecture. The metadata details recorded include the original object, the digital object, digitization process, technical details and staffing details. The records are created by the digitizer and overseen by one person. This metadata record for the digital deliverables is then held in a separate catalog in electronic form on an intranet server and is available on the internet.

The relation of the records for the digital deliverables to those for the original digitized materials is a mixture of independent and identical. A text string links the catalog and the object.

28.5 Format, Resolution and Compression of Digitized Materials

The format of retroconverted text is SGML and XML. This text contains non-Latin scripts and OCR (Prime Recognition) is used to convert the digital images. The accuracy of this process is generally high but variable, at best 99.95%. Specific treatment to pages prior to OCR is to clean page images. The aim of using OCR was to allow automatic indexing, enhanced searching and computer-based analysis of the materials. The project's recommendations on OCR are to use high-end software and, if money is available, to outsource it. No keying in is used in-house.

The TIFF file format is used for capturing and preserving. Some capture is done in JPEG, but this is mainly the delivery format. 20-40MB per image is the typical file size at between 8 and 24 bit-depth. Lossless compression is used at the capture and delivery stages when low-resolution cameras are in use. The project undertakes many kinds of post-processing using PhotoShop. The recommendation from the project's experience of digitizing images is to produce several versions at high resolutions. Moving image digitization used videodisc format.

The quality control procedures are spot tests on texts, and spot and total checks on books. Metadata are double-checked. The project quality control predicates the whole workflow process, and highlights the need to train staff.

Users do not have to pay to use the digital deliverables, rather they have access to them through an open access catalog (although some material is not available). Glimpse is used for full text searching/browsing and sequel and Postgress queries are used on the metadata. Users are also able to zoom in on images.

Potential users of the digital deliverables are informed about their availability through website announcements, articles in print media, conferences and meetings and email shots. The project does not know which medium has been the most effective but it is probably building links to the site. Usage levels are 300,000 page impressions in 24 hours, with 30,000 users, as monitored by automatic data capture.

28.6 Evaluation, Funding and Long-term Sustainability

The project has carried out front-end, formative and summative evaluations that included questionnaires on paper and online, focus group discussions and observing users' interactions. See:
http://ils.unc.edu/~march/perseus/lib-trends-final.pdf

An estimate of the project funding cost is $10 million; $2.5 million was for the Classics Computer Project (CCP), $6 million from Federal Agencies, mainly the National Endowment for the Humanities (NEH). With regard to an ideal cost, the project would estimate $6-7 million, if the project were now to be repeated. The project would not have done anything differently. The project believes that the use of standards has saved money and that to invest in these in the short-term will have long-term benefits. The funding agencies monitor the project through Annual Reports, but the most significant sanction is continued funding.

New material (and its associated metadata) is updated on an ongoing basis. The user interface is changed intermittently and file formats, especially the back-end data, are changed incrementally.

The project has a preservation strategy (the university is creating a digital depository) as well as SGML to XML

and different storage media based on a migration data strategy.

The digital deliverables will be available indefinitely and the project's longer-term sustainability will rely on self-generating funds in future. The project has resources for the next four years. The exit strategy for the project is through the university depository. Loss of the digital deliverable would be a matter of concern.

valid xhtml 1.1
abp~04/02

29 Library of Virginia (LVA)

HATII interviewed Elizabeth Roderick, Director of the Digital Library Program at the Library of Virginia, on September 26 2000. The Program was initiated in 1995, by the Library's Collection Management Services department, in an effort to preserve, digitize and enable access to important archival and library collections. Through digitization, the materials held by the Library are therefore integrated into an accessible research environment. The project is responsible for the digitization of 2.2 million original documents, as well as the creation of eighty searchable databases, indexes and finding aids. Items of historical and cultural significance can thereby reach a wider audience, both in terms of educational usage and public access.

29.1 Organizational Digitization Program and Policy

The Library of Virginia's (LVA) digitization strategy developed alongside the growth of the web, the availability of web-based interfaces to bibliographic databases, and the development of the MARC 856 field. These developments coalesced around 1994/95 with LVA pilot projects, available finances and capital for equipment.

The project conducted an informal collection survey that interviewed staff members and convened *ad hoc* groups from within the library to brainstorm about the collections. This survey involved people from all branches of the Library and developed a shortlist of projects in the Library and archives. The biggest obstacle to overcome in carrying out the collection survey was the uncertainty that it prompted. If the project were to conduct a survey now it would make it more formalized, and would disseminate information more widely (because some staff felt excluded or did not understand the objectives).

The collection survey was used to establish priorities for digitization. The project director, who had worked towards a consensus and was free of political considerations, established these priorities in consultation with appropriate managers. The priorities were not formalized in a strategic policy statement but practicability was the biggest factor involved. The project director will develop a formalized policy in the next twelve months.

One of the objectives this policy sought to achieve was to establish some sort of institutional memory - a framework that would ensure that all staff would work towards a common goal. Another objective was one of validation for the Digital Library Program itself, in that it should be consultative and not *ad hoc*. The project believes that it has been successful in meeting these objectives. The recommendations the project would make to other organizations that are attempting to formalize their selection criteria are: to make sure there is one person in ultimate charge; to be clear about the institution's mission and its primary customers; to gain support from higher administration; to make detailed documentation; to talk to collection experts; and to minimize open-ended or unresolved questions that will produce repercussions.

Obstacles to planning the development and building of the digital deliverables have included staff members who were uncomfortable with the project and were slow to support it, along with apprehension on the part of the library's small technology staff at the time.

The criteria that have guided the project in selecting and prioritizing materials for digitization are: intellectual property rights (IPR); teaching and learning potential; conservation; historical and cultural value; research significance; provision of user services; enhanced access; research into digitization strategies; preservation; labor cost reduction; potential to reach disadvantaged groups; and infrastructure cost reductions. Of these, meeting user needs, especially on high use materials, has been top priority. The project is not a preservation strategy; material is microfilmed first and nothing is discarded. This saves cost in terms of delivery. The priority is then practical feasibility (such as housing and storage conditions). These criteria and their priority have not changed over time.

The project has co-operated with archives, libraries, museums, academic institutions, corporations, foundations and charities, and government agencies at the institutional, local, regional and national levels. Program staff regularly provides consulting services to these institutions. Co-operation with national bodies such as the Library of Congress has been on specific projects (e.g. religious petitions and national newspapers, which the LC

digitized). Recommendations from this experience are to be clear about the approach, and to determine if or when it is appropriate or desirable for institutions other than the holding Library to digitize and/or host collections. The project would also suggest digitizing materials in their entirety, so as not to create artificial collections.

The current status of the program is ongoing; it started in 1994/95 with no anticipated end date.

The primary purposes in creating the digital deliverables are public access, provision of teaching and learning resource, wider access and response to previous demand. Secondary purposes are preservation and research. The program has produced an explicit statement of intent that covers its rationale, scope and significance.

The type of source material digitized includes:

- printed documents;
- unbound printed documents;
- typescript documents;
- hand-written documents;
- hand copies;
- photographs.

The format of the materials to be digitized covers a wide range up to large-scale maps (which could be delivered using MrSID, via LizardTech.). In some cases secondary surrogates needed to be made (e.g. colonial records photocopied). Digitization is intended to represent the entire body of material.

The following standards, guidelines or tools are used for representing content:

- Unicode;
- JPEG;
- MARC.

The following standards, guidelines or tools are used for describing content:

- MARC;
- Dublin Core.

The following standards, guidelines or tools are used for controlling data values:

- AAT;
- LC subject headings;
- LC Thesaurus for Graphic Materials;
- name authority files.

The program looked at other existing guidelines for digitizing particular document types when planning its digitization strategy. In particular the program's vendor (VTLS Inc., Blacksburg, Virginia) was quite visionary and the program relied on its expertise, especially on early solutions for graphical material. Other early pioneers were also consulted to achieve the simplest, highest possible quality, non-proprietary master.

The following standards, guidelines or tools are used for representing structure:

- HTML;
- Unicode;
- JPEG.

In relation to standards in general and suggestions for navigating between the ideal and the realistic, the program would recommend making things as simple as possible by planning file names, structures, and common elements at the beginning, and also to make decisions about standards early, thereby reducing the burden on data entry staff.

The primary intended audiences for digital deliverables are K-12, community college, four-year college, graduate school, lifelong learning, distance learning, museum users, public library users, archive users, government and private sector. The project has not carried out any evaluation of the target audiences. Groups other than the primary target audiences could use the digital deliverables.

The program has not acknowledged the needs of those with disabilities or the W3C's "Guidelines for Web Site Accessibility". There are no limitations to the use of the digital deliverables because the program has steered away from IPR issues and has attempted to operate in good faith in all cases. The profile of users has been as

anticipated, as the aim was to make heavily used material available in digital form. However, the program has been surprised at the level of customer support requests (people with different browsers, errors, explaining collections so that expectations are managed, etc.). The program has since instituted an error tracking system.

29.2 Project Management and Planning

There is considerable internal experience in project management in the person of the Director. The program was established as part of the Library (Collection Management Services Division). The program has led to changes in organizational relationships and procedures. It was appropriate for the program to become part of the Collection Management Services (CMS) Division because the latter houses technical services, archives processing and description, preservation services and special collections. One project management procedure that has not been successful, has been coordination with the Library's IT Department. Most IT tasks are handled by the vendor, and on many occasions local Library IT staff have not been available to provide necessary on-site support for the Program. Quality assurance is achieved through the use of staff with a background and interest in the subject matter of the collections, and expertise and knowledge of cataloging and imaging, rather than computer expertise

Regarding pilot and feasibility studies, the program undertakes sampling and testing on a case-by-case basis for technical feasibility, workflow analysis and workflow piloting. As a result of these studies, some projects have been scrapped as priorities have shifted. Cost benefit, time and motion and benchmarking studies have also been carried out. Work is assigned by the program's three project managers who sequence and schedule each project. They then assign to part-time staff or outsource.

Approximately fifty percent of the work is done on-site at the LVA and the remaining fifty percent is completed by the vendor, particularly large-scale batch data processing projects. The factors in each decision are the resources available, technical expertise and what material can be moved. The program has a track record of five years with the vendor. The program purchased equipment, including flatbed scanners, drum scanners, film scanners, digital cameras and high-end professional cameras. Guidelines have been established for data capture procedures and benchmarks used are grayscales and color charts.

29.3 Human Resources and Training

The project employs four full-time (classified) staff, one director (100%), three project managers (100%) and ten PT, 40 hours-per-week (non-classified) staff to digitize, catalog, analyze and re-house materials. Staff have a predominantly library background or content/subject expertise and none was redeployed from other areas. External advice was sought on the technical aspects of digitization through reading, discussion, workshops and self-study.

Identifying training needs was not a problem, but finding quality training was. Areas where training needs were identified are preparation and handling of materials for digitization, technical operation of digitization equipment, post-digitization and metadata creation/cataloging. The project director, full-time and part-time staff engaged in training and this was organized in-house (via internal and external consultants), external courses, independent study and learning on the job. The training has so far met the needs of the project.

29.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and owns the copyright of the original materials. The program is working on a declaration of copyright or rights status to protect the property. No third party copyright material is digitized. Users are allowed to make printouts of the digital deliverables on paper. Users can view HTML, and compressed TIFF IV, JPEG and MrSID images. No electronic management systems, such as watermarking, are in use.

The project has a conservation procedure for the original materials whose condition is investigated by appropriate staff, such as archival or special collections staff. If necessary, the program carries out conservation procedures such as rebinding or mounting. The project modifies, degrades or compromises some material to carry out digitization (e.g. dis-binding, sometimes the original). Risk assessment of the material during the preparation for digitization is undertaken on an informal basis. Special equipment is available but the program does not digitize books. Materials are prepared by curatorial or preservation staff prior to digitization but are not

monitored by them during digitization. Once the material has been digitized, restrictions are placed on the originals as a preventative measure to reduce handling and serving pressure.

Various cataloging and reference systems were in place prior to digitization, ranging from no finding aids or old finding aids, to index cards and databases. All of the available information from the catalog or reference systems is used in the digitization process, and staff members conduct extensive research to identify all relevant information and to locate core reference or source material. The program does alter some originals for the digitization process. Material has been rejected for incoherent pagination, incomplete works and blurred pages. The program has digitized from originals and intermediaries and some material has only existed in intermediary form. Intermediaries that have been used are photocopies, slides/35mm or 4x5 transparencies, photographic prints, microfilm and single frame microfiche.

The original material catalogs range from handlists to databases. The library and archive catalogs are both on the web and there are 35 different bibliographic databases. Cataloging is done at the item level wherever possible with bibliographic records linked to individual images via the MARC 856 field. MARC and USMARC standards are used in cataloging digital deliverables. The program emphasizes the bibliographic description of the original item and provides only brief information about the digitized image in the 533 field. However, the digitization process and image format is standard for each type of material, is consistent across collections, and is documented at the program level.

Tools for controlling data values are LC subject headings and the Art and Architecture thesaurus and LC thesaurus for graphic materials. Details recorded in the metadata include information about the original object and the digital object. The vendor also maps from a variety of catalog sources to MARC records (using MARC Maker). The digitizer and an archivist/information professional record metadata, and some skeletal MARC records are used. Scanning and cataloging are separate in the workflow but the filename is captured when scanning (this is set up prior to scanning). Metadata records for the digital deliverables are then held in separate bibliographic databases from the main online book catalog, which is available on the Internet. The database records and the digital objects are linked through the MARC 856 field. In the case of photographs, there is a one-to-one correspondence between the 856 field and the image; in the case of multi-page documents, the 856 field is linked to an intermediate HTML page that contains links to the individual document pages.

29.5 Format, Resolution and Compression of Digitized Materials.

Format for retroconverted text-based digital deliverables includes:

- HTML (very few sources, probably only three);
- Database (MARC).

None of the texts contained non-Latin scripts. OCR is used by the vendor but only very little on one or two projects (pre-published, typescript material, e.g. the published guide to the Board of Public Works collection). The aim of using OCR was to achieve searchable computer-based analysis. From this experience the key advice is to use clean text.

Keying-in and data-mapping are the main conversion methods for textual materials. The lesson learned here concerned data entry errors created by the vendor's part-time staff due to inadequate supervision, which later had to be manually corrected.

The TIFF file format is used for capturing, preserving and delivery. Photo CD is also a capture and preservation format. JPEG is solely a delivery format.

The capture and preservation resolution is 300dpi (600 capture for small images and then back to 300 for archiving). Delivery resolution is between 72 and 100dpi. Bit-depth is 24 throughout. Image compression is CTIT TIFFs for preserving and delivery and JPEG for delivery. For the *Camp Lee Soldiers Home Applications and Registers* a grayscale JPEG presentation was also used. The aim of compression was to improve access, enhance usability, decrease storage requirements and reduce the risk of theft of the digital resources. The project retains the uncompressed scans.

The program carries out post processing on the images (and sometimes text) as well as rescanning from the originals. This is done when there is a poor horizontal or vertical crop. De-skew, re-size, grayscale to bitmap and rotation (from microfilm) are carried out.

Average file size for capture and preservation is 1MB (6MB for 8x10 photos), compressed Group4 images are

1MB and up to 100K uncompressed. The dynamic range of the equipment is checked through the set up calibration. From their experience of digitizing images the program recommends planning, analyzing and going through the raw material and similar materials at the same time, looking at the physical organization of the resource to allow for re-housing/handling and attempting the best capture from the outset.

The quality control procedures in place for the digital deliverables are spot checks on map records (subject-heading keywords can be checked easily). Total checks are undertaken on photograph and document images. Link checking is accomplished with spreadsheets. Metadata quality control procedures involve a double check (e.g. Bible records are under one accession and then by alphabetic character). An ASCII text file is then produced which the vendor place checks. Scanned images and then a burned CD image are also checked for legibility.

Users can access the open catalog plus materials. Searching and browsing facilities are offered through the Library's integrated system (through MARC records). Facilities include keyword, Boolean and combination searches. Browsing is approached through the catalog, and the second stage is thumbnails. When there is no MARC or HTML information, there is a card index-like record with links. Apart from searching and browsing users are able to manipulate images by zooming, rotating, etc., using MrSID. Users must employ a TIFF viewer as a browser helper application in order to view the TIFF images. The Library provides a generic TIFF viewer at no cost, and users may also obtain other viewers such as Netmanager's Chameleon and Netview, Wang image, Alternatiff, and GraphicConverter for MAC. Usage is monitored by automatic data capture and formal data collection. Usage for a typical month will include users from more than 130 countries, with 120,000 hits on all product Home Pages, 45,000 database searches performed, 147,000 digital images and 1 million electronic cards viewed.

Users do not have to pay for the use of the digital deliverables, but have the opportunity to order high quality images off-line.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, print and broadcast media coverage, conferences and meetings, email shots, registering with search engines, receptions and word of mouth. Many users are gained through search engines, word of mouth and list servers for target groups (e.g. Virginia and genealogy lists, the Scout Report).

29.6 Evaluation, Funding and Long-term Sustainability

The program carried out front-end evaluation of users before the development of the project. The techniques used were online questionnaires, email and focus group discussions. As a result of this valuable feedback, the program modified and enhanced several products, sometimes in several versions, to better serve the needs of the users. Formative evaluation with users was achieved through online questionnaires, email and focus group discussions. The results from this were to enhance the interface, navigation and amend errors. Summative evaluation will be undertaken with the Virginia Historical Inventory Project (Mellon funded). In addition, all feedback information is retained.

The lesson from this evaluation is simple: users do not read instructions and putting material online can lead to unforeseen problems, particularly with regard to remote user support.

Since 1995 the program has cost $4 million plus salaries. Funding was from federal and grant sources. The program believes that its $4 million funding is sufficient and has actually asked to have its budget reduced. The program believes that the use of standards and guidelines has saved money in the short and long term. The program has been monitored by its funders through annual reports to the Institute of Museum and Library Services (IMLS), Mellon and internal reporting. Mellon requested cost models and workflow reports from the program.

New material (and its associated metadata), metadata change, user interface updates and file format changes are ongoing but it was not known at what frequency.

The program's preservation strategy is offsite storage and secure backup. Text and image materials have storage media and storage condition strategies, while images also have a file format strategy. These strategies are based on migration of data.

The digital deliverables will be available indefinitely and the project's longer-term sustainability is not reliant on self-generating funds. The project has secured funding at a national level. The project does not have an exit

strategy, but may start charging if this became necessary. Loss of the digital deliverable would be a matter of concern.

valid xhtml 1.0 strict
abp~04/02

30 University of Virginia, Center for Digital History (Alderman Memorial Library)

On September 25 2000, HATII interviewed William G. Thomas, Director of the Virginia Center for Digital History at the University of Virginia. The project at the Center for Digital History was initiated as a study of American history and culture, in order to subsequently transform teaching methods in this field. The new technologies were employed with a view to advancing the discipline of historical study. To this end, the project produced high quality, reliable resources with World Wide Web availability and therefore, the capacity to reach wider audiences, both within the education community and in the general public.

30.1 Organizational Digitization Program and Policy

The most important criteria for selecting and prioritizing material for digitizing are teaching and learning potential and research significance. Medium level criteria are historical and cultural value, enhanced access, potential to reach disadvantaged groups and social inclusion. The selection and prioritization criteria have not been changed over time.

Co-operation in developing the Center's digitization program has involved libraries (UVA Special Collections), museums (Woodrow Wilson Birthplace), academic institutions (Virginia Tech, Norfolk State University), corporations (Public TV), foundations and charities (National Council for History Education) and government agencies (Department of Education — State).

A recommendation from this experience is that it is essential to build strong collaborations across disciplines and with schools both within and outside the institution.

The program was initiated in 1991/92 when IATH started work, and the Center was established in its own right in 1998. There is no anticipated end date.

The purposes for which the digital deliverables are created and their rank are:

- teaching and learning (1);
- research (2);
- public access (3);
- experiment (4);
- preservation (5 — possibly)

The project produced an explicit statement of intent that covered its rationale, scope, significance and primary audience.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed illustrations;
- hand-written documents;
- artifacts;
- intermediaries;
- photographs;
- sound recordings;
- video recordings;
- other (large maps).

The nature and format of paper material is relatively standard and, at the other extreme, maps are up to 6ft by 6ft in size.

Where possible, the program generally tries to represent an entire body. Some of the digital deliverables are intended to be re-purposed, e.g. Valley of the Shadow CD-ROM. This approach has had mixed results: the images have been reasonable but the maps and databases were challenging, and nothing was seamless. The problems are inherent in the discrepancy between the tools that are available for CD versus those available for the web.

Standards, guidelines and tools used for representing content are:

- TEI;
- JPEG;
- SGML;
- XML.

Standards, guidelines and tools used for describing content are:

- TEI header;
- Other (XML DTD for newspapers).

No standards, guidelines or tools are used for controlling data values, as none were appropriate for the project.

The Center consulted some existing guidelines for digitizing particular document types, possibly in the Making of America White Paper.

Standards, guidelines and tools used for representing structure are:

- TEI;
- HTML;
- JPEG;
- SGML;
- XML.

A recommendation for how to navigate between ideal standards and realistic use, is that while this is a function of funding and staff ability, it is nevertheless important to aim for the highest quality at every level.

The highest priority intended audiences are K-12, community college, four-year college and graduate school. Lifelong learning, distance learning, museum users, public library users, archive users, government, and the private sector rate medium priority.

An evaluation of the target audience occurred through a 1996 NEH seminar for teachers at UVA. This was not in a social scientific way but through workshops and seminars; a report was posted on e-text. Groups other than the target audience can use the digital deliverables. The Center is 75% compliant with W3C's guidelines for web accessibility, for example, in terms of text browser capability.

Some projects have restrictions on the use of the digital deliverables because of copyright. These restrictions are clearly stated.

30.2 Project Management and Planning

Both external and in-house project management advice was available.

The Center for Digital History is related to the structure of UVA within the college of arts and sciences as a Research Center, and therefore has no curriculum.

The digitization program has led to the university changing its allocation of resources; the very existence of the Center is proof of this policy change. There are no formal project management procedures in place, but ideally every project will have a manager and director/directors. It has been noted that the decision to eliminate the post of project manager was a mistake. On distributed projects, site managers are also required, otherwise communication breaks down.

A large number of quality assurance procedures are in place, including style sheets for markup and data entry and a team leader in each group to provide day-to-day leadership, so that the manager is not solely responsible. In addition, a second opinion is required, concerning all material.

Pilot studies were carried out for training needs but the program underestimated technical needs. As a result the program raised more money. No formal benchmark or time and motion studies have been undertaken. Gantt charts are used to aid project management. Work is delegated according to the interest and specialization of students. Job descriptions and performance indicators are used (this is a state requirement).

Digitization is a combination of in-house and outsourcing (some programming, web design, newspaper images for Valley project for the latter). Outsourcing is chosen on a project and material basis and where the skills are not available in-house. Equipment for in-house digitization was bought in through a large university grant (ITC). When the material excludes some digitization processes, the choice is then based on quality assurance and lowest error rate.

The following technologies are used:

- flatbed scanners;
- film scanners;
- digital cameras;
- high-end professional cameras.

Guidelines for data capture procedures have been established. Benchmarks are used mainly for data and text.

30.3 Human Resources and Training

The Center employs one director (100%), one associate director (also metadata specialist, currently vacant), five digitizers/research assistants (10 hours per week), one technical development worker (50%), one educational specialist (25%) and one administrator (33%). Staff are predominantly from an arts, humanities or history background. The administrative support was redeployed from another area.

Advice on the technical aspects of digitization was available both externally (through manufacturer's support) and in-house.

Training needs were assessed informally and areas identified were:

- Post-digitization processes;
- metadata creation.

Specialist technical staffs were engaged in training in-house, with internal consultants. The training has met the needs of the project.

30.4 Project Life-cycle Processes and Procedures

The Center is aware of the copyright position of the digital deliverables. It does not own the copyright in the original materials (which for the most part are out of copyright). The copyright status is declared in an overall statement, which has been effective. Material in copyright is digitized with the owners' agreement, payment of a fee or without formalities. There are no copying restrictions for users.

Users are able to view and download text:

- ASCII;
- TEI DTD marked-up text;
- XML marked-up text;
- PDF.

Users can view and download images:

- Thumbnails;
- lower quality images;
- highest quality images;
- associated documents.

Audio resources are listen-only through full length compressed (streaming) sound.

Moving images are view-only through lower and highest quality digital video clips.

No electronic management system is in use.

The Center does not have a conservation procedure for the original material and little investigation into the condition of the material is made prior to digitization. Conservation activities are left in the hands of the collection holder. To carry out digitization one set of newspapers was dis-bound. No risk to the materials was identified during the preparation or digitization process. No special processes or equipment are specified but curatorial staffs prepare material prior to digitization. There are no access restrictions on materials after digitization.

It is not known which cataloging and referencing systems are in place before digitization. Some information from these records is adopted. The Center has to locate some core reference material. Some material is altered from the original to complete the digitization process. No material is rejected. The Center uses reproductions or intermediaries where the material only existed in intermediary form.

The form of reproductions/intermediaries used includes:

- photographic prints;
- microfilm.

The digital deliverables are cataloged as part of project management but the records are not transferred to the library.

Standards or guidelines used for cataloging the digital deliverables are not applicable. An in-house system using MS Access is used to catalog digital material, which is later expressed in the TEI header.

Tools used for controlling data values are:

- controlled vocabulary (in-house).

Metadata details recorded are:

- original object;
- digital object;
- staffing details.

The digitizer creates Metadata records. The records are on an intranet server and are available on the internet. The records for the digital deliverable and the original digitized materials are independent of each other. The catalog and object are linked through the database.

30.5 Format, Resolution and Compression of Digitized Materials

The format of retroconverted texts includes:

- HTML;
- SGML (TEI light);
- XML (own DTDs).

No texts contained non-Latin scripts.

No OCR software is used. Keying-in is used as the method for textual materials. The lesson from this is to use style guides for documents.

Capture and preservation formats for images are TIFF and delivery formats are GIF and JPEG. Capture and preservation resolution is between 300 and 600dpi and delivery resolution is 72 or 100 to 150. Bit-depth is not known. JPEG compression is used for delivery to reduce cost, improve access and decrease storage requirements. The Center retains the uncompressed scans. The program does not carry out any processing on the images.

An outside consultant carried out sound digitization for the Presidential Tapes Project.

Capture formats are:

- WAV and Ra.

Preservation format is:

- Ra.

Delivery format is:

- Ra.

Quality control for the digital deliverables includes spot checks on likely error areas and total checks on a second pass. Metadata quality control procedures are second pass checks. Quality control procedures have meant the workflow has been prolonged.

Users have a combination of open access to the catalog, and restricted access to the materials, open access to which is limited to in-house users. For web delivery, users are able to search by keyword, Boolean, etc. (including metadata). The Center's specialist software uses are Generator, Flash Generator and MrSID. Users are able to zoom in on MrSID images.

The level of use of the digital deliverables is 13,000 hits a day monitored by automatic data capture using Webusage 7.0. Users do not have to pay to use the digital deliverables.

Potential users of the digital deliverables are informed by website announcements, press releases, articles in print media, print and broadcast media coverage, conferences, meetings, email shots and registering with web search engines. Print and broadcast coverage has proved the most effective.

30.6 Evaluation, Funding and Long-term Sustainability

The Center's formative evaluation used the following techniques:

- focus group discussions;
- user observation;
- computer interaction logging.

As a result of this, evaluation changes were made to the interface and nature of materials.

The program has cost $1,500,000 so far (not including institutional support). The main sources of funding have been UVA, NEH and private donors. The main problem with the funding has been that it has been unevenly distributed. Therefore, if the Center were given funding again it would prefer a level of base support and then project funding. The Center believes that adopting standards represents a cost because higher expertise is needed.

Funding organizations monitor the Center through annual reports (plus quarterly reports for UVA/NEH). Funding organizations required project management plans, cost models and workflow reports from the project.

New materials and metadata are added daily and the user interface is changed every six months.

The Center does not have a preservation strategy beyond CD-ROM (server material is backed up by ITC). The Center intends to keep the digital deliverables available indefinitely. The long-term sustainability of the program is dependent on securing future grants.

There are a variety of options for the Center's exit strategy e.g. commercial opportunities. Loss of the digital deliverable would be a matter of concern.

30.7 Additional Comments

The interviewee felt the questionnaire had a digital library focus.

31 University of Virginia, Electronic Text Center

David Seaman, Director of the Electronic Text Center at the University of Virginia, was interviewed by HATII on September 22 2000. The Center began work in 1992, with two intentions: to construct and maintain a collection of SGML texts and images that are accessible on the Internet, and to create user communities capable of navigating the collection and handling the materials. The project would ideally reach audiences outside the University community, therefore allowing the digitized material to fulfil its potential to extend the user demographic.

31.1 Organizational Digitization Program and Policy

The Electronic Text Center did not conduct a collection survey of its own. However, firm collection management policies were already in place within the library and the project bought in what material they could at the start in 1992. Since then there has been a growing preference from vendors and publishers for titles in electronic form with standard markup. This matches the Center's original aims of providing an online archive of standards-based texts and images and building a user community to create and use them.

Although the Center did not have its own collection survey, its early priorities for digitization were the high-use materials requested by students from the library, such as Twain and, latterly, Japanese texts. Inaccessible items, such as those within Special Collections were also a priority. Those with an interest in a project, mainly the collection specialist, E-Text Center staff and academic users, establish the priorities for digitization. These priorities have not been formalized in a policy statement but the Center is beginning to look at digitization for preservation and undertaking digitization for reserve readings in the library.

The primary objective of the Center's digitization policy is to serve multiple audiences and build collections for this. Within this multiple audience Virginia academic users and other partners have first priority, followed by all other users. The Center feels it has been successful in achieving these objectives as usage continues to grow and the site statistics are watched closely to see which texts people are using.

Recommendations for other organizations that are attempting to formalize selection criteria in strategic policy statements are not to take on too much — one or two projects with a well-chosen sub-set focus is most appropriate. These must show relevance, especially to the core institution. Organizations must also play to their home strengths and emphasize content and context.

The largest overall obstacle for planning the development of digital deliverables was securing funding. Co-ordinating the project partners was also a challenge, as was preserving constituents' sense of ownership.

Research, pedagogy and access drive the whole selection and prioritization of materials to digitize. However, there is an increasing movement towards digitizing material for preservation. Consequently, the decision to digitize material is dictated by physical condition and is less dependent on content. The biggest change over time has been in scale, for example the use of vendors.

The Electronic Text Center's co-operation has involved archives, libraries, museums, academic institutions, corporations, foundations and charities, and government agencies, as well as high schools and individuals. This co-operation has been across all levels — institutional, local, regional, national and international.

The Center began in 1992 and the current status of the program is ongoing with no anticipated end date.

The primary purposes for which the digital deliverables are created, are as a teaching and learning resource, as research material, and to provide public and wider access. Secondary purposes are preservation, experiment and revenue generation.

The project produced an explicit statement of intent that covered its rationale, scope, significance and primary audience in a brochure and on the web.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- printed illustrations;
- handwritten documents;
- artifacts;
- photographs.

There are no very large format materials that are being digitized by the Center and the material is exclusively paper-based.

Inevitably the material that is digitized is a sample but where possible, the Center tries to digitize all the material within a collection. The material is intended to be re-purposed for things such as course packs, e-books and print-on-demand (which is just being introduced by organizations such as Barnes and Noble booksellers). The aim is to generate revenue in a non-evasive way. No one expects to get print books for free, therefore some form of micro-charging, for example, 10c for an e-book, generates revenue while the value remains greater than the charge.

Standards, guidelines and tools used for representing content are:

- TEI;
- Unicode;
- JPEG;
- SGML;
- XML.

Standards, guidelines and tools used for describing content are:

- MARC;
- Dublin Core;
- EAD;
- TEI Header.

Standards, guidelines and tools used for controlling data values are:

- LC subject headings;
- Name authority files.

As guidelines, the Center looked at TEI and EAD for digitizing particular document types.

Standards, guidelines and tools used for representing structure are:

- TEI;
- HTML;
- Unicode;
- JPEG;
- SGML;
- XML.

The Electronic Text Center perceives a nested series of audiences. Of these UVA staff and students have first priority (along with four-year college and graduate school), but the general public also assume a high priority by their sheer bulk, alongside community college and K-12 audiences. A third priority is the growing home schooling movement. These audiences have not been the ones anticipated due to the speed of web development and its rapid accessibility outside the academic community. Furthermore, the Center believes that projects need to be more explicit about what they can do, for example, reaching high school audiences. Evaluation of the target audience occurs through two annual library surveys. The project has not taken account of the W3C's "Guidelines for Web Accessibility".

There are limitations on the use of the digital deliverables because of copyright restrictions, for example the state-wide Viva consortium for purchasing materials does not allow the E-Text Center to re-sell, claim ownership or to mirror elsewhere.

31.2 Project Management and Planning

Because of the early start of the Center, it has tended towards involvement of consultants regarding the management of digitization. The E-Text Center's management is firmly part of the University of Virginia's library structure.

As with many University organizations the UVA library is a rigidly hierarchical organization and the Center exists because one senior librarian was prepared to take the risks. There is now an increasing willingness to assign staff and faculty positions (which is how the E-Text Center got underway).

There are no formal management structures in place and the greatest luxury for the Center has been to be able to act like a start-up in the critical early years and work at the speed of a business. For quality assurance, the Center's Director meets weekly for half an hour with his immediate supervisor. In addition, there are several issue-based committees within the library.

Neither pilot nor feasibility studies were carried out at the start because the project was small-scale and "ramping up". The Center has not carried out time and motion studies or benchmarking except where a grant required it, but it recognizes that this is not ideal and not necessarily an exemplary model.

Job descriptions are highly formalized and semi-annual performance evaluations are undertaken. The Center's work relies significantly on graduate students, however, it also has a growing full-time staff base. Each person has responsibility within the Center; for example a graduate student is appointed as the point person for a project.

Bulk digitization is outsourced for reasons of cost-effectiveness, however, smaller volumes and conversion are carried out in-house. Initially there was no in-house equipment except for cameras, and no significant funding was available for equipment, certainly not enough to go large-scale and ensure value.

High-end professional cameras (two Phase One cameras 10,000 x 12,000 pixels) are used for color digitization of rare books.

Data capture procedures were established for the above mentioned cameras and associated equipment (database, CD burning, etc. – see EAF Scanning Procedures Document for further details), which have quality assurance as the primary objective.

The benchmark used for image digitization is Kodak color strip for each book.

31.3 Human Resources and Training

The Center employs one FT director, an associate director and an assistant director, a project manager and a programmer. In addition, there are approximately twelve graduate students during semester at 10 hours per week, rising to 30 hours during the summer. With the exception of the programmer all have a humanities background and have risen through the ranks of the project.

Advice on the technical aspects of digitization was available in-house.

Training has largely been undertaken in-house and expertise developed at the same rate as the Center itself. Most of the Center staff has had some bibliographical training prior to this job, which has been invaluable. At the beginning, all staff had to learn new computing methods and programs, but their previous use of computers in teaching and research provided a strong foundation for learning new systems. The staff members' strong research and pedagogical backgrounds have also given the Center a good understanding of what a humanities researcher and teacher may want from such a system. Close contacts with UVA departments also helps to lure in new graduate student and faculty users.

Worthy of note, are the Center's efforts in training its users. This training (for example in HTML or TEI) has helped build the user community to the extent that users can create their own texts and contribute them to the Center's collection.

31.4 Project Life-Cycle Processes and Procedures

The Center is aware of the copyright position of the digital deliverables. It does not own the copyright in all the

original materials. The copyright status is declared.

Depending on copyright and access restrictions, users are able to download TEI and XML texts.

No electronic management systems are in use, other than domain restrictions.

The Center does not have a conservation procedure for the original material, except insofar as it is beginning to digitize material for preservation.

The original materials are already cataloged by their providers. Some digital surrogates are cataloged on the library's OPAC system, which is also used for the originals.

Standards or guidelines used for cataloging the digital deliverables are:

- Dublin Core (within some projects);
- MARC (derived from TEI/DTD).

Tools used for controlling data values are:

- controlled vocabulary (LC files)
- classification scheme (local, e.g. for the Civil War project to build browser lists).

Metadata details recorded are:

- original object;
- digital object;
- digitization process;
- technical details (to a degree);
- staffing details;
- administrative information.

Metadata records are created by an archivist/information professional, while a library cataloger fixes errors. The metadata records are included in the main (library) catalog, which is in electronic form and available on the internet. The relationship between the records for the digital deliverable and the original digitized materials is the same. The catalog and object are linked by the 856 field URL link.

31.5 Format, Resolution and Compression of Digitized Materials

The format for retroconverted text material is:

- SGML;
- XML.

Some texts contained non-Latin scripts (e.g. Japanese).

OCR is used for smaller jobs and to clean modern typefaces. However, for older materials and when working with large-scale materials, it is more appropriate to use keying-in, as it is the best, fastest and most inexpensive method.

For images the capture and preservation format is TIFF and the delivery format is JPEG.

Capture and preservation resolution is between 400 and 600dpi and delivery resolution 100dpi. Bit- depth throughout is 24-bit color and JPEG compression is used for delivery to improve access. The program retains the uncompressed scans. The program carries out processing on the JPEG files using PhotoShop but does not alter the original TIFF files. The average 100 dpi jpeg file size is 100K, while the 24-bit 400 dpi tiff original is around 30-40 megabytes. The dynamic range of the equipment is calibrated and recorded in the metadata.

A recommendation from this area would be that when an object can be scanned as well on a flatbed scanner as on a digital camera, then it would be preferable to use the flatbed scanner (for example in the case of a smaller manuscript page). The Center provides a high level of public access to their browsable and searchable texts. Thousands of these texts and images are publicly accessible, although these etexts are not necessarily in the public domain (they cannot be re-published without the Center's permission).

Because of contractual obligations with the vendors who supply the texts and the search software, access to

parts of the on-line text service is restricted to the parties covered by the various license agreements (usually University of Virginia users only).

31.6 Evaluation, Funding and Long-term Sustainability

The Center has not conducted any formal evaluations, but the close contact and involvement of users means the Center enjoys increased feedback on all aspects of their work: collection development, text encoding, image type and quality, and the design of the search and browse pages. A central lesson has been the validation of some of the Center's primary choices: the 24-bit TIFF color images and the TEI texts that have produced for years – somewhat on the promise of future software and hardware developments – are providing the flexibility the Center needs to respond to a changing web environment and a developing set of user demands.

Users are able to browse and search by keyword according to collection, subject, publisher, place of publication and further by author's name or date range. Compound searches can also be performed.

Users do not have to pay to use the digital deliverables, but non-UVA users cannot access material that has been licensed to just to UVA.

Potential users are informed about the digital deliverables by website announcements, press releases, articles in print media, print media coverage, conferences and meetings and email shots. The most effective dissemination strategy, in terms of building a sustainable resource, is based on the Center's local focus, and the grounding effect this produces.

New works are added on a continuing basis, and the Center is working their way through the public ASCII texts on the Internet whose provenance can be determined.

Each day the Center receives around 38,000 individual user hits from 19,000 unique internet hosts machines, accessing over 130,000 items. This is monitored by automatic data capture.

valid xhtml 1.0 strict
abp~04/02

32 University of Virginia, Institute for Advanced Technology in the Humanities (IATH)

HATII interviewed John Unsworth, Director of the Institute for Advanced Technology in the Humanities at the University of Virginia, on September 11 2000. IATH is a research unit of the University, which aims to expand the employment of IT within a humanities environment. In overseeing the digitization activities of the University, IATH is responsible for producing digital materials with specific relevance to research interests. While maintaining this focus on the activities of the University, the Institute also forms partnerships and participates in initiatives with other institutions, in order to further the development of IT in the field of cultural heritage.

32.1 Organizational Digitization Program and Policy

The Institute for Advanced Technology in the Humanities (IATH) does not have an organizational digitization strategy *per se*. Any strategy is decided case by case for the various projects that IATH runs. These strategies are based on an analysis of the materials; size of image and processing issues are a factor in any strategy. The documentation associated with the projects will include the digitization strategy. Much of IATH's advice to the projects that are considering establishing a digitization strategy is based on the realities of the materials and of the processes. IATH does not conduct a collection survey, as each project is different and the collections are not necessarily all local, in fact very are often not. While many institutions that control the artifacts will often carry out the digitization, this is not always the case with IATH. Each project will differ; some will carry out digitization in-house on locally controlled materials, while other projects will send out to other institutions for the digitization process. The selection of the materials is driven by research rather than by the institutional collections.

As the digitization projects are research-driven, there is often a rush to production and a pressure to produce the materials for research. These obstacles can often work against a solid digitization design process. If the project has succumbed to this pressure, then building the digital deliverables can be a more difficult process; often there is a lack of training in faculty staff. A more formal structure may remedy this.

Selection of materials is highly research-based with intellectual property rights, teaching and learning potential, historical and cultural value, enhanced access, research into digitization strategies, potential to reach disadvantaged groups and social inclusion all as important factors. The selection criteria have not changed over time, as research is the main impetus.

IATH has collaborated with archives, libraries, museums, academic institutions, corporations, foundations and charities, and government agencies at the institutional and national levels. This collaboration has been very successful, with the main criteria being to establish firmly who has responsibility for which task.

IATH has many projects, which have differing purposes, priorities, and objectives. Many projects are complete and as many again are ongoing. Generally, projects do not finish entirely and have some level of addition. Priorities for projects will include preservation, public access, teaching and learning resource, research, wider access and response to previous demand, as well as specific criteria such as urban planning (this would be for a very specific purpose).

Projects do tend to produce websites with explicit statements of intent covering rationale, scope and significance.

As the projects cover a wide range of topics, the type of source material digitized can include:

- printed books, serials, and similar bound materials;
- printed documents;
- unbound printed documents;
- typescript documents;
- printed illustrations;
- handwritten documents;

- artifacts;
- original artworks;
- intermediaries;
- mechanical copies;
- hand copies;
- photographs;
- textiles;
- film;
- sound recordings;
- video recordings;
- GIS data;
- topography.

The format of the materials to be digitized covers a wide range as specified by the project criteria and will include items such as drawings and photographs, schematic plans of water systems and large-scale maps. Depending on the project, materials may either be a representative sample or the entire collection.

Different projects have different views on the process and what may be changed, for example understanding digitization of maps, and using different specialist software.

Again, the range of projects will cover the use of the following standards, guidelines or tools for representing content:

- TEI;
- Unicode;
- JPEG;
- MARC;
- SGML;
- SQL;
- Flash.

The following standards, guidelines or tools have been consulted for describing content:

- MARC;
- Dublin Core;
- EAD;
- TEI Header;
- CWA.

Projects will use a database schema that refers to these standards.

The following standards, guidelines or tools were consulted for controlling data values:

- LC subject headings;
- LC subject thesaurus;
- Name authority files.

Different projects use standards in different ways, and will, for example, consult them for creating their own database schema.

When standards are not used, it can be for a variety of reasons. They may not yet be agreed, they may be inappropriate or not known, or there may be a lack of documentation or of training.

IATH projects will consult other existing guidelines for digitizing particular document types when planning their digitization strategy; an example would be standards for large-scale maps.

Projects use different standards, guidelines or tools for representing structure, including:

- TEI;
- HTML;
- JPEG;
- MPEG;
- SGML;
- XML.

In relation to standards in general and recommendations for striking a balance between the ideal and the realistic, the program recommends starting with the standards and constructing their own database which can have better data ranges. However, it is advisable to use and consult standards wherever possible.

The range of projects ensures that the primary intended audience for digital deliverables will cover many communities such as K-12, community college, four-year college, graduate school, lifelong learning, distance learning, museum users, archive users, government and private sector. The main audience is research users, generally in the graduate school community. Some projects have carried out evaluation of the target audiences using methods such as focus groups of the main institutions and it would be expected that a wider scope of researchers would use the digital deliverables.

The program has not acknowledged the needs of those with disabilities through the W3C's "Guidelines for Web Site Accessibility". Some of the materials will have limited access for IPR reasons. It is not known if the profile of the users differs from the anticipated set.

32.2 Project Management and Planning

IATH uses in-house expertise when planning a digital project and talks to other centers at a technical level. IATH tends to be the funder of the projects and the project managers are recruited by Faculty. IATH will give advice and training where required and each project will differ in the reporting relationship but it tends to be very informal. Some projects have more formal procedures such as selection committees and annual reviews.

One project management procedure that has not worked is the flat project where the structure is very democratic. It was difficult to establish intellectual responsibility. IATH encourages projects to develop the project management procedures best suited to individual needs. Quality assurance procedures will depend on the project varying from a little to extremely high standards.

Just as each project varies on procedures, they also vary on planning, and occasionally pilot and feasibility studies will be undertaken. However, no project uses benchmarking for scheduling tasks. Planning tools will vary from project to project: IATH will encourage but not prescribe.

Digitization can be carried out in-house or outsourced depending on the project. The factors in each decision are the resources available, and permission for access. IATH has equipment already installed. These include flatbed scanners, film scanners, digital cameras and high-end professional cameras. Guidelines for data capture procedures have been established and benchmarks used are gray scales and color charts but, once again, this will vary from project to project.

32.3 Human Resources and Training

Each project will have a variety of staff but IATH has three directors, one metadata specialist, four technical support/development staff and two administration support staff. However, it is feasible that a hundred students may also be employed. Forty to fifty faculty staff may be involved in the various projects.

Staff have a wide variety backgrounds including library background or computing science as well as architecture and literature publishing. Initially, staff were redeployed from other areas. Internal and external advice was sought on the technical aspects of digitization.

Areas where training needs were identified are project management, applications of technical standards, preparation and handling of materials for digitization, technical operation of digitization equipment, post digitization and metadata and markup. The project director, specialist technical staff, library cataloging staff, equipment operators and, in fact, everyone on the projects are all engaged in training and this has been organized in-house (via own consultants), external courses, independent study and learning on the job from each other. It is considered that training can never fully meet the needs of the projects as staff can never have enough training in this developing area.

32.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and the ownership of the copyright of the

original materials will depend on the project. The projects declare the copyright on the page as text embedded in the object, for example as JPEG headers. This practice has proved effective. Some material in copyright is digitized and this is done under legal provisions for libraries, with the owner's permission and by payment of a fee.

Users are allowed to make printouts of the digital deliverables on paper, download to a PC, LAN and WAN. Users can view and download text, image audio and video objects in different ways on a project-to-project basis. Potentially these will be: ASCII text files, TEI DTD marked-up text; thumbnails, lower quality images, highest quality images, associated documents; full length compressed sound and lower quality video clips. Watermarking is deliberately not used, as IATH believes that it defaces the original and misleads the users.

IATH does not have a conservation procedure for the original materials, as this is not applicable to the organization. The condition is investigated by appropriate staff, such as archival or special collections staff, but any conservation would be carried out by the institution in possession of the material. Risk assessment of the material during the preparation for digitization is undertaken by analysis. Objects are handled carefully and gloves are worn. Some projects will use special equipment. Potentially as each project varies, materials can be prepared by curatorial or preservation staff prior to digitization and may be monitored by them during digitization. Once the material has been digitized, restrictions can be placed on the originals as a preventative measure to reduce handling and serving pressure depending on the project.

As IATH does not own the material the cataloging and reference systems will vary depending on the institution that the materials came from. Projects often have to use the information from these systems but often do not have access to all the relevant information and have to locate some core reference or source material. It is possible that some originals are altered for the digitization process.

Various projects have been digitized from originals and intermediaries and some material has only existed in intermediary form. Intermediaries used include photocopies, slides/35mm or 4x5 transparencies, photographic prints, microfilm, single frame microfiche and glass negative. Audio intermediaries include 78rpm discs, 33 rpm discs, audio cassettes and digital sound media. Moving images can be digested from video copies, betacam SP and 16mm.

The original materials are not cataloged by IATH. The digital surrogates are cataloged for internal project management systems. They have used Astoria document management and Unix and relational databases. Tools consulted for controlling data values are controlled vocabulary, thesaurus (Getty) and classification schemes. There is no off-shelf solution for the projects. Details recorded in the metadata are information about the original object, the digital object, the digital process, technical details, staffing details and administrative information. The creation of metadata is project dependent.

32.5 Format, Resolution and Compression of Digitized Materials

Format for retroconverted text-based digital deliverables varies from project to project and can include:

- HTML markup;
- SGML markup;
- XML markup;
- Database;
- spreadsheet (occasionally);
- Unicode.

Some of the texts may contain non-Latin scripts. Keying-in rather than OCR is used to convert the digital image. From this experience the key advice is that it is more cost-effective to send out large quantities of text and is worthwhile to combine projects for this purpose.

The TIFF file format is used for capturing and preserving; JPEG is solely a delivery format. JPEG 2000, Wavelet and MrSID are also used for delivery.

The capture and preservation resolution and bit-depth will depend on the project but capture is at the highest resolution appropriate to the subject. Images are compressed using Wavelet, JPEG and MrSID. This is done to improve access and decrease storage. The originals are retained in uncompressed format.

The program carries out post-processing on the materials using PhotoShop; the processing techniques varying from project to project. Due to the wide range of projects there is no average size of file. The dynamic range of

the equipment is checked through the set-up calibration.

Sound files are digitized at the digital media center. WAV, AIF and RealAudio are used for capture and delivery of sound. The sampling and bit-rate will depend on the project. Projects have used bit-rate compression and frequency compression but once again the details depend on the project. Compression is used to improve access and to decrease storage. Sound files will have noise cleaned in post-processing procedures.

Moving images are also digitized by sending them to the digital media center. File formats used are AVI, MPEG, QuickTime, RealVideo and Mov for capture and delivery. The resolution (frames per second) will depend on the project. Kodak video compression is the current favorite and this compression is used to improve access and to decrease storage. Video is post-processed to clean up noise and edit to bite-sized chunks.

IATH does not have a formal quality control procedure in place for all projects and each project will have its own. IATH has oversight of the projects and advises on these procedures. Rendering in markup is one method to ensure metadata quality control.

Users can access the materials generally on the web and access tends not to be restricted. Search facilities include DynaWeb, indexing HTML, Excite search engine and PostgreSQL. Users can search the metadata using DynaWeb and databases. No special hardware or software is required. Users can resize objects online. Each project varies on the level of use but these are recorded through web statistics project by project.

Users do not have to pay for the use of the digital deliverables for any project.

Potential users of the digital deliverables are informed about their availability through website announcements, articles in print media, print and broadcast media coverage, email shots, conventional mail shots, registering with search engines, receptions and bibliographic studies. Many users are gained through email and bibliographic studies.

32.6 Evaluation, Funding and Long-term Sustainability

IATH does not have formal evaluation procedures for the projects and any front-end evaluation is done in the selection process of the materials by speaking to researchers. In these cases comment forms are used. Some aspects of projects have been changed as a result of front-end evaluation, such as assessment of the time to start the project.

No formative evaluation is carried out and no project has really ended, so there is no summative evaluation.

IATH has funded approximately six to seven million dollars worth of projects. Half of this funding comes from the university and the other half from private funds, federal funding and corporate donation.

New material (and its associated metadata), metadata change, and user interface updates are ongoing at various frequencies depending on the project. File formats are not changed if possible.

The variety of projects means a variety of approaches to preservation. However, each project will ensure that the strategy used is appropriate to the materials and the goals of the project.

IATH concentrates on the scholarly goals for each project and the scholar is the driver for all projects. They recommend using appropriate standards where possible. Projects should look at the goals and why they are using documents, and create their digital strategy based on these criteria.

valid xhtml 1.0 strict
abp~04/02

33 University of Virginia, Robertson Media Center (Clemons Library)

On September 22 2000, HATII interviewed Judith Thomas, Assistant Director of the Robertson Media Center at the University of Virginia. Since its inception in 1991, the Center has sought to provide individual and group access to a broad range of materials at the University. The collection includes more than 17 000 videos and audio recordings, images and artifacts in various formats. While the primary target audience is the University community, for teaching and learning purposes, the digitization process inevitably enables wider access to the materials to those outside the university. Throughout the past year, the Center has initiated several developments to its practices, most significantly the increased centralization of digital activities across all units of the library. The recent creation of the Digital Library Production Service to act as a gatekeeper of the digital collections, while also being the locus of all production activity, is also intended to promote co-operation and adherence to standards.

33.1 Organizational Digitization Program and Policy

The Robertson Media Center's digitization strategy was initiated and is driven by curricular requirements, such as the early demands of the Art and Archaeology departments in the early 1990s to develop a response to the use of proprietary tools for images. A further impetus was given by the move to the web via non-searchable study guides. Audio material was developed for out-of-class study, foreign languages and special collections' presentations. Video material was developed for media- centric curricula.

A collection survey was undertaken including faculty and libraries, but this was a very project-orientated exercise and cannot be scaled for the whole of the population. The Center's activities began in an entrepreneurial way however, it has recently changed this approach. The Center has been making advances in the areas of collection development, project management and digital practices, with a view to increasing the centralization of digital activities, with collaboration across all units. With these incentives in mind, the Center developed a system for managing the central repository of digital resources. This work is overseen by the Digital Library Research Group, led by Thornton Staples.

Priorities for digitization come through separate processes in a project. Batch is used wherever possible, followed by the post-digitization processing, cursory description and metadata creation.

The objective of the Center's digitization policy is to fulfil faculty and students' curricular needs. The Center's collection efforts were initially solely in response to curricular needs, but the newly instigated collection development strategy for digital images allows for a systematic building of the collection.

Overall obstacles for planning the development of digital deliverables were the lack of institutional support for tools that the Center has identified as critical. Obstacles for building digital deliverables were that collection building for non-text material was difficult as there were no institutional solutions, or the solutions were elusive.

The primary criteria for selecting and prioritizing material for digitizing are: teaching and learning potential (including historical and cultural value), and enhanced access. In the case of material digitized for special collections, conservation is a priority. Secondary criteria are research significance and preservation. The quality of the original and IPR are further considerations.

The selection criteria have not changed over time, but prioritization criteria have done so. There has been an increased interest in preservation, especially of motion media because of technological advances.

Co-operation has involved archives, libraries, museums and government agencies. This co-operation has been mostly local, at the regional level on the question of how material could be shared, at the national level with MESL and at the international level with UNESCO.

One lesson was the lack of compatible systems at the regional level in order to share resources.

The Center began in 1991/92 and the current status of the program is ongoing with no anticipated end date.

The first purposes for which the digital deliverables are created, are as a teaching and learning resource and to provide wider access. Secondary purposes are research and preservation.

The project produced an explicit statement of intent that covered its rationale, scope, significance and primary audience for fundraising and has had much iteration on the web.

The type of source material digitized includes:

- printed illustrations;
- artifacts;
- original artworks;
- photographs;
- textiles;
- film;
- sound recordings;
- video recordings;
- radio broadcasts.

The nature and format of the material digitized includes:

- image (direct capture, printed material, slides and negatives);
- video (laser discs, VHS, NTSC, PAL, DVD and film);
- audio (CD, cassette, open reel, direct capture).

Whether the digital deliverables represent a sample or an entire body of material depends on the collection; for example, if it fits other selection or prioritization criteria for the whole of the collection. It is surprisingly seldom that the digital deliverables are intended to be re-purposed, for example as an educational CD-ROM.

Standards, guidelines and tools used for representing content are:

- CIMI (Dublin Core DTD);
- JPEG;
- MPEG;
- XML;
- HyTime (under consideration);
- TIFF;
- AVI;
- Other audio formats (Real, AIF).

Standards, guidelines and tools used for describing content are:

- CIMI (Dublin Core DTD);
- MARC;
- Dublin Core;
- Categories for the description of Works of Art
- Other (VRA)

Standards, guidelines and tools used for controlling data values are:

- AAT;
- LC subject headings;
- LC subject thesaurus;
- Name authority files;
- Other (CONCLASS).

The guidelines which the Center consulted for digitizing particular document types were emanating from Getty, MESL, Harvard, Michigan and Oxford (Stuart Lee). The Center looked at these to corroborate its own experience and they were modified for local requirements.

The MARC standard was rejected because it did not work for digital images, but the Center maps its records to MARC.

Standards, guidelines and tools used for representing structure are:

- CIMI (Dublin Core DTD);
- HTML;
- JPEG;
- MPEG;
- XML;
- Other (Georeferred information, FLASH).

In terms of navigation between ideal standards and realistic use, the Center's recommendation is to interpret locally then make sure it can map for interoperability.

The first priority target audience is four-year college and graduate school students.

An evaluation of the target audience was made through faculty meetings, library content surveys and identified needs.

The 30% of the digital deliverables that are openly accessible can be used by groups other than the target audience, such as other sections of higher education, as well as there being great interest from K-12. The profile of users has been as anticipated. The project has not taken account of the needs of those with disabilities or the W3Cs web accessibility guidelines.

There are restrictions on the use of the digital deliverables because of the original rights and copyright holders, and these restrictions are clearly stated.

33.2 Project Management and Planning

The Center follows homegrown project management guidelines.

The Robertson Media Center is related to the structure of UVA as a library department, but collaborates with computing. The digitization program has led to the broadening of political boundaries and tighter collaboration across centers. The work of the Digital Library Research Group and the creation of the Digital Library Production Service have encouraged these developments.

Quality assurance procedures are benchmarking for different types of procedures, such as references and questions for traditional library matters, but few for digital projects, since so many of these are open and fluid. Concept testing is in place and smaller projects are storyboarded.

Pilot studies for scheduling, training needs, technical feasibility, user needs, workflow analysis and piloting and technology are used to assist project management and planning. Significant changes were made to the program as a result of these pilot studies in the areas of technology, training and workflow.

The electronic tools used for scheduling are flow charts and diagrams. Clearly defined areas of technical expertise determine who does what work. Public service and project support is shared. Job descriptions and performance indicators are used.

Most digitization is outsourced for reasons of quantity, cost effectiveness and terms of grants, which make it economical to outsource. Digitization equipment was bought in and the decision as to which digitization process to adopt is based on media type and the requirement for making an archival and/or endurable copy.

The following technologies are used:

- flatbed scanners (Epson, HP, Microtec and Pixelcraft);
- film scanners (Kodak, Micon and Epson);
- digital cameras (Kodak);
- digital video (Canon);
- audio (a variety of equipment).

Guidelines for data capture procedures are flexible, and are increasingly adhered to, since the development of a more controlled environment. Benchmarks are a minimal color calibration.

33.3 Human Resources and Training

The Center employs one full-time director (also as metadata specialist) and five full-time technical specialists (who also carry out digitization work). Each member of staff also fulfils a technical support and development role (approximately 20% of their time). In addition between one and five graduate students are employed for 5-10 hours a week. The project also works closely with staff in the school of education. Most staff have a humanities background and were employed from other areas.

Advice on the technical aspects of digitization was available externally and in-house. Areas where training was or is needed were identified as:

- application of technical standards;
- preparation and handling of materials for digitization;
- technical operation of digitizing equipment;
- post digitization processes;
- metadata creation;
- digital preservation (selective, e.g. audio).

The Center also looks at the training needs of the end users; for example, media experts teach graduate students. All staff were or are engaged in training in-house, with internal consultants and through learning on the job. The training has met the needs of the project.

33.4 Project Life-Cycle Processes and Procedures

The Center is aware of the copyright position of the digital deliverables in some cases. It does not own the copyright in the original materials (which are retained by the original donors or holders). The copyright status is declared next to the acquisition device simply as “refer to copyright law”. Material in copyright is digitized under the legal provision for libraries and the Center offers copyright advice to anyone who seeks it.

Depending on the copyright position, users are able to download thumbnails, lower and higher quality images and associated documents. Most audio material is available for listening only in less than 30-second samples, full length compressed, high fidelity sound and associated documents. For digital moving image most material is available for viewing only as samples, lower quality clips, highest quality clips and associated documents.

IP access restrictions are in place.

The Center does not have a conservation procedure for the original material.

A rough assessment of the originals is made by the Center, before being referred to experts. The Center seldom undertakes conservation activities on original materials except for audio. No material is modified, destroyed, degraded or compromised to carry out digitization. Only small subsets of fragile originals are identified as being at risk during preparation for digitization.

As playback is destructive for audio and video, risk is assessed through feeding the materials. To minimize risk, fragile material moves back to special collections, and vulnerable materials are then made stable for digitization. Conservation staffs sometimes prepare material before digitization but do not monitor it during the process. Some access restrictions are placed on materials after digitization.

Cataloging and referencing systems in place before digitization include BPAC and hand lists, and as much information is taken from these as possible. The Center usually has access to all the relevant catalog materials but has to locate some core reference material. Some material is altered from the original to complete the digitization process and some material is rejected because of blurred pages, or for reasons of general quality or fragility. In many instances the Center uses reproductions or intermediaries and in these cases the material only existed in intermediary form.

The form of reproductions/intermediaries used includes the following.

For images:

- slides, 35mm or 4x5 transparencies;
- photographic prints;
- microfilm;

single frame microfiche.

For audio:

- 78 rpm discs;
- rpm discs;
- cassettes;
- digital sound media.

For moving images:

- video copies from film;
- betacam SP.

The digital surrogates are often cataloged differently from the originals.

Standards or guidelines used for cataloging the digital deliverables are:

- Dublin Core;
- MARC;
- USMARC;
- VRA;
- Other (CDWA).

Tools used for controlling data values are:

- controlled vocabulary (AAT, ULAN, Iconclass and LC subject terms);
- thesaurus (LC subject terms)
- classification scheme (in-house, under development by digital library).

Metadata details recorded are:

- original object;
- digital object;
- digitization process;
- technical details;
- staffing details;
- administrative information.

Metadata records are created by the digitizer, archivist or information professional and by the digitizer who is also archivist or information professional. The metadata records are included in the main catalog and a separate catalog which is in electronic form on an intranet and available on the internet. The records for the digital deliverable and the original digitized materials share some information. The catalog and object are linked by the filename reference.

33.5 Format, Resolution and Compression of Digitized Materials

For textual material the formats used are:

- HTML;
- XML (refer to digital library for unique object identifier);
- database (filename).

Some texts contain non-Latin scripts.

The OCR software used is Omnipage and 80% accuracy was achieved with no special treatment carried out beforehand. The aim of using OCR was to capture machine-readable text for user based needs. Keying-in is not used as a method for textual materials.

For images the capture formats used are GIF, TIFF, Photo-CD and PDF. Preservation format is TIFF and delivery format is JPEG, PDF and MrSID. Capture and preservation resolution is 2000-3000 pixels one side and screen resolution for delivery. The bit-depth for illustrations, rare books and manuscripts is 24-bit color. In the case of microfilm or delicate material, the preferred quality setting is 1-bit 600 DPI tiff files, with 8-bit 400 DPI tiff

files used on occasion, for materials such as photographs or handwritten documents on microfilm. LZW compression is used at capture and JPEG for delivery. The aim of compression is to improve access, enhance usability and decrease storage requirements. The program retains the uncompressed scans. The program processes JPEG images using PhotoShop and Debalelizer. MrSID images have their own processing engine. The average file size at capture and preservation is 25MB, under 1 MB for delivery. The dynamic range of the equipment is not checked.

Recommendations from this area would be to acquire familiarity with issues of color fidelity and capture at high resolutions.

Methods used for the digitization of sound are analog play, mixing, digital capture, and manipulation and compressing.

Capture and preservation formats are AIF, MIDI, Au and QuickTime.

Delivery formats are AIF, Real Audio, MIDI, Ra, Au, MPEG and QuickTime.

Compression is used and the aim is to improve access, enhance usability and decrease storage requirements. Post-editing is carried out.

For moving images the capture and preservation formats are QT and MOV. Delivery formats are MPEG-1, QT, Real Video and MOV. Moving image resolution for capture, preservation and delivery is NTSC 30 fps. Capture and preservation compression is DV and for delivery MPEG and QT. The aim of compression is to improve access, enhance usability and decrease storage requirements. Post-capture processing includes adding data, fades, and altering frame rate and size using Media Cleaner, FinalcutPro and Imovie.

Quality control for the digital deliverables includes spot checks, a random set of checks and checks on a stratified random sample. Metadata quality control procedures involve checking by the project leader. Quality control procedures have influenced the Center's assessment of time and manpower.

Users do not have to pay to use the digital deliverables. Access is a combination of: open access to the catalog, the catalog and materials; access restricted to in-house users; and restricted access. For web delivery, users are able to browse at the collection level and view individual items. Metadata information is delivered with the digital object (retrieved from database). The program has no special hardware or software needs. Users are not able to manipulate any of the digital deliverables except for MrSID images.

The level of use of the digital deliverables is primarily course support at the moment. Usage is monitored by automatic data capture using Netwatcher and Webstats.

Potential users of the digital deliverables are informed by website announcements, press releases, articles in print media, conferences, meetings, email shots, conventional mail shots and registering with web search engines. It is not known which of these has proved the most effective.

33.6 Evaluation, Funding and Long-term Sustainability

The Center has not yet undertaken any evaluations.

It is not known how much the program has cost so far. The main sources of funding have been grants and library budget allocations. The Center believes that the use of standards has been costly in the short-term but has produced savings in the long-term. Funding organizations monitor the Center through periodic reports and assessment of deliverables.

New materials and metadata are added with high frequency.

A preservation strategy is under development. Quality control procedures are in place for life-cycle management according to media type. The Center intends to keep the digital deliverables available indefinitely.

The Center does not have an exit strategy. Loss of the digital deliverable would be a matter of concern.

valid xhtml 1.0 strict

34 University of Virginia, Special Collections (Alderman Memorial Library)

Edward Gaynor, Associate Director of the Special Collections department at the University of Virginia, was interviewed by HATII on September 22 2000. With a particular interest in American history and literature, the department at the Alderman Memorial Library had the task of digitizing material including more than 12 million manuscripts. The material represents primary and secondary historical sources, which can subsequently be used to support teaching and research. The material to be digitized is subject to evaluation of its fragility, prior to inclusion in the project.

34.1 Organizational Digitization Program and Policy

Priorities for digitization were based on the strengths of special collections; for example, the ability to secure grants is based on the appeal of certain collections, such as African American photographs. These priorities were not formalized in a strategic policy statement. The first objective of digitization is to serve user needs and, secondly, to increase access to collections; it is not a preservation strategy. The project has been successful in achieving these objectives.

The main obstacle to developing and building the digital deliverables was the absence of an overall plan — independent centers developed by format. Efforts to integrate all formats into one digital delivery system are underway.

The primary selection criteria are the provision of user services and enhanced access. Other considerations are research significance and teaching and learning potential. However, Special Collections does not evaluate the importance of the user's need for the material. As far as increased access is concerned, the material is either heavily used or else it is an important collection that is little used. These criteria are fairly constant although Special Collections is now somewhat less constrained about what it delivers.

The project has co-operated significantly within the library but not much outside the university, with the exception of libraries in North Carolina. One recommendation from their experience is that high expectations of other departments, which are time-dependent, may not materialize.

The project began in 1994 and is ongoing. It has no anticipated end date.

The first purpose of creating the digital deliverables is as a teaching and learning resource and, secondly, to provide wider access. Experimentation was also an early purpose but this was never explicit. Revenue generation is a minor factor, mainly to cut out frivolous requests (there is no charge for UVA users external users have to pay).

The project produced an explicit statement of intent covering its significance.

The nature and source of materials digitized are:

- printed books, serials and similar bound material;
- printed documents;
- unbound printed documents;
- typescript documents;
- printed illustrations;
- hand-written documents;
- artifacts;
- original artworks;
- intermediaries;
- mechanical copies;
- hand copies;
- photographs.

There is a huge range of materials right up to large maps; the only material that is relatively standard is nineteenth century manuscripts. The digital deliverables are neither a sample nor the entire collection. Each collection is evaluated but usually it comprises the material that the user has requested. The project would like to re-purpose material but since it lacks the staff to do so itself, the material is provided for others to re-purpose.

- The standards, guidelines or tools used for representing content are:
- TEI;
- CIMI (Dublin Core STD);
- JPEG;
- SGML;
- XML.

The standards, guidelines or tools used for describing content are:

- MARC;
- Dublin Core;
- EAD;
- TEI Header.

The standards, guidelines or tools used for controlling data values are:

- LC subject headings;
- LC subject thesaurus;
- name authority files.

The project did not consult guidelines for digitizing particular document types.

The standards, guidelines or tools used for representing structure are:

- TEI;
- CIMI;
- HTML;
- JPEG;
- SGML;
- XML.

In relation to the use of standards in general and how to use them in practice, the project believes in interoperability as far as possible, using standards where feasible, but not worrying about detailed implementation. Implementation of library-wide digital standards will change this approach.

The intended audience for the digital deliverables and their priority are:

- K-12 (low priority);
- community college (high priority);
- four-year college (high priority);
- graduate school (high priority);
- lifelong learning (low priority);
- museum users (low priority);
- public library users (low priority);
- archive users (high priority).

The project has not undertaken any evaluation of the target audience, and the materials could be used by those outwith the target audience. The project originally looked at the needs of those with disabilities and the W3C's guidelines on web accessibility.

There are no restrictions on materials placed on the web. Other material is restricted because of limitations on the originals, such as copyright and restrictions to UVA users only. The profile of users is as anticipated but online exhibits have attracted a higher proportion of K-12 users.

34.2 Project Management and Planning

Informal internal and external advice was taken on project management. The project started as quite separate

from the rest of the library and UVA but over the last year or two they have tried to communicate developments widely and to expand internal projects. There are no formal project management procedures in place but some are developing library-wide. One project management procedure that has failed, has been the lack of a single area in charge of an entire project.

Quality assurance procedure is random spot-checking. This developed from the first project where quality assurance was omitted and two-thirds proved incorrect.

Neither feasibility nor pilot studies have been carried out. Several time studies have been conducted. This started with the EAF (Early American Fiction) project and is now a standard part of projects. The parameters used are to lay out the work, break it down to the smallest part and extrapolate from these. It is the quantifiable part of this process that has proved the hardest to measure, for example, network transfer rate.

Planning and scheduling tools are becoming more elaborate — calendar schedules tied to a tracking database to generate quality assurance and backups. It was not difficult to delegate, since people were hired specifically for projects. Non-project staff include Michelle Kraft who is a digitizer and manages the student digitizers. Both job descriptions and performance indicators are state requirements.

Digitization is carried out in-house for cost reasons. Equipment was bought in for digitization, some for each project, and Special Collections was the first to receive equipment. The choice of digitization process is sometimes equipment-based; they have some expertise but lack the facilities for large plan digitization.

The following technologies are used:

- flatbed scanners (HP, Umax);
- one digital camera;
- two Phase Ones and one Kontron high-end digital cameras, plus a 10 year old Nikon slide scanner.

Guidelines are established for data capture procedures and benchmarks in use are gray scales, color and reproduction charts.

34.3 Human Resources and Training

Personnel employed on the project include the following.

Special Collections Staff:

- one director (25%);
- one digitizer (100%).

Grant Project Staff:

- five FTE;
- 2.5 FTE;
- between one and six student digitization assistants (0.75 FTE, 10 hours a week).

Most staff have a humanities background, one has a library background. Staff were hired for their skills or subject background, and areas where training was needed were:

- project management;
- application of technical standards;
- preparation and handling;
- technical operation;
- post digitization processes;
- metadata creation.

The project director, specialist technical staff and library or cataloging staff all engaged in training. This was organized in-house (with their own consultants), through external courses, independent study and learning on the job, and has met the requirements of the project.

34.4 Project Life-Cycle Processes and Procedures

The project is aware of its copyright position regarding the digital deliverables, but does not own the copyright in the originals. This is retained by donors and in most cases resides with the IPR holder (e.g. literature) and digitization has been carried out under the legal provision for libraries. The copyright status of the deliverables is stated on the website but not on each item. This practice has been effective.

Users can download material for individual use. They can view and download TEI, DTD and XML marked-up text; and download thumbnails, low quality and highest quality images, and associated documents. There are no electronic management systems such as watermarking in use to control copyright.

There is always a full conservation survey carried out on each item prior to digitization and the project will not digitize materials which are fragile or may be damaged by the process. No material is modified, degraded or compromised to carry out digitization. There was no official risk assessment during preparation or digitization but new cool fluorescent and UV filtered lights have been installed to replace those which were halogen and non-UV filtered. Other steps to minimize risk are the use of bookrests and cradles. Curatorial staff demonstrate how to handle material during digitization and monitor the process. There are no official restrictions on the originals after digitization, but users are first directed towards the surrogate. In some cases, use of the originals has increased.

Card or online catalogs, printed guides, encoded guides are in place before digitization and basic descriptive or Dublin Core elements are used in digitization. The project does not have to locate any core reference material. Material is rejected for digitization because of skewing and fragility. Sometimes the project uses intermediaries, but the material does not just exist in this form. Forms of intermediaries used are slides, 35mm or 4x5 transparencies, photographic prints and microfilm.

The project does not catalog every surrogate, but when it does it uses the same method as the original. Standards used for this are:

- Dublin Core;
- EAD;
- TEI Header;
- USMARC.

Filemaker pro is used for tracking; cataloging at the item level is not merged. Tools for controlling data values are:

- controlled vocabulary (Library of Congress, and own);
- thesaurus (Library of Congress subject headings).

Metadata details recorded are:

- the original object;
- the digital object;
- administrative information.

Some metadata records for the digital object are included in the main catalog and some are not. The catalog and the objects are linked by the 856 field.

34.5 Format, Resolution and Compression of Digitized Materials

The format for retroconverted text is:

- HTML;
- SGML (TEI/EAD);
- XML;
- database (Filemaker, Access);
- spreadsheet (Excel).

None of the text contained non-Latin script.

OCR is used on rare occasions (Omnipage Pro). The accuracy level depends on the text and no special treatment is carried out beforehand. Their advice is not to OCR any material pre-dating 1850.

All manuscript material is keyed-in and their advice here is to use operators who are familiar with the subject area and can read copperplate script.

Images are preserved and delivered in GIF, captured, preserved and delivered in TIFF and JPEG, and delivered as PDF.

Capture and preservation resolution is 600dpi, delivery 72-300dpi, bit-depth for all is 24-bit color. JPEG compression is used for delivery, firstly to improve access, and secondly to decrease storage requirements. The project retains the uncompressed originals. Color balance, gamma, re-sizing and compressing are carried out using PhotoShop. The dynamic range of the equipment is checked.

Quality assurance procedures are random spot checks and on rare occasions total checks. Such quality assurance procedures have an impact on workflow and planning for it in advance is recommended.

Users have open access to the catalog and about 75% of the materials. Searching and browsing facilities are open text, images through Filemaker (metadata in Filemaker). The level of usage on the last collection was 200 hits a day or 10,000 a month, measured by automatic data capture.

Users do not pay to access the deliverables.

Potential users are informed by:

- website;
- press releases;
- articles in print media;
- print and broadcast media coverage;
- conferences, meetings;
- email shots;
- conventional mail shots;
- registering with search engines.

The most effective of these are email and direct mail.

34.6 Evaluation, Funding and Long-term Sustainability

No front-end evaluation was undertaken. Formative evaluation used paper questionnaires, email and user observation, and no significant changes were made as a result of this. Summative evaluation uses paper and online questionnaires, email and user observation. The purpose of this was to find out whether surrogates online were an adequate substitute for originals or whether they increased original use. These results were not disseminated.

The project has cost $1,750,000 to $1,800,000, 90% of which was from grants (Federal Agencies, NEH, IMLS, Mellon). The project feels it should have cost $2 million. If it had had the extra money, the project would have made more rapid progress. So far the quality has been good, but it has taken a long time. In the long-term standards have saved money, but have possibly cost money in the short-term.

Funding organizations monitor the project through reports (one quarterly, one annual, Mellon semi-annual plus site visits).

Documents required by funders are:

- collection survey;
- conservation plan;
- project management plan;
- cost models;
- workflow reports;
- economic cost/benefit analyses.

New material and metadata are constantly added. Metadata are enhanced yearly, as is the user interface. It is not clear when file formats are changed.

The project is committed to keeping the materials viable and available indefinitely, but has not determined how. Loss of the deliverables would be a matter of concern.

valid xhtml 1.0 strict

abp~04/02

35 University of Virginia, William Blake Archive

HATII interviewed Matthew Kirschenbaum, the Technical Editor of the William Blake Archive at the University of Virginia, and Andrea Laue, the Project Manager, on September 22 2000. Sponsored by the Library of Congress, the William Blake Archive represents a digital collection of work by and relating to Blake, in various forms from printed books to artifacts. Rather than restricting access to the University community, the Archive aims primarily to reach an international audience. Indeed, this was the most significant motivation for its initiation. Matthew Kirschenbaum completed the questionnaire section of the interview instrument, while Andrea Laue was interviewed for the remainder.

35.1 Organizational Digitization Program and Policy

The William Blake Archive developed in response to the burgeoning interest in Blake's poetry and visual art amongst a dispersed and varied global audience. The aim of the Archive is to provide an international, unified, public resource. Active since 1996, the Archive has sought to unite visual and literary works that are highly disparate, widely dispersed, and often severely restricted because of their value, rarity, and fragility.

The William Blake Archive's priorities for digitization were formalized in the original Getty Grant Program application, produced by the project editors. In the first phase (pre-1999) this prioritized a group of Blake's illuminated books based on historical principles derived from recent revisionist scholarship. These works were also the best known and studied by the editors; however, gaining permission from institutions has meant that presented difficulties from the outset. In the second phase (1999 to 2002), priority has moved to additional illuminated books and non-illuminated material, such as paintings, drawings, original and reproductive prints, manuscripts, and rare or unique typographical works. The aim behind this second phase is to expand the core of the Archive and exemplify the full range of Blake's work. Ultimately it is hoped that the full range of Blake's work could be incorporated.

The objectives this policy sought to achieve are to broaden access, enhance teaching and scholarly research through providing comparison copies and editions, and provide a durable foundation for future scholarship and a legacy of adaptable tools.

The biggest obstacle to planning the development and building of the digital deliverables has been obtaining permission (which are the largest single budgetary expense), along with the actual cost of the digitization process.

The criteria that have guided the project in selecting material are those that informed the original editorial principles. These principles have not changed over time, even after the shift to non-illustrated books as the corpus of illustrated books became digitized.

The project has co-operated with libraries, museums and academic institutions at the national and international levels. In particular the Archive hopes that the collaborative work of the three editors and the staff of the Institute for Advanced Technology in the Humanities will become a useful prototype of "distance editing".

The Archive's work began in 1995 when the Getty Grant Program underwrote the initial phase of the project and has an anticipated end date of 2004.

The purposes in creating the digital deliverables are academic and public access, teaching, learning, research, experiment, and in response to previous demand. The program has produced an explicit statement of intent that covers its rationale, scope, significance, primary audience, long-term sustainability, level of faithfulness to the originals and suitability for different target audiences. This statement is available on the web at http://jefferson.village.virginia.edu/blake/public/about/

The type of source material digitized includes:

printed books, serials and similar bound material;
- unbound printed documents;
- printed Illustrations;
- hand-written documents;
- artifacts;
- original artworks.

It is hoped that the digitization process will ultimately represent the entire body of material. The Archive is being built on the principle of creating a history of Blake's artistic production in one edition. However, the first priority was to digitize one copy from every printing of every book, with the illuminated books as an editorial and archival backbone. The project does not intend to re-purpose these digital deliverables.

The following standards, guidelines or tools are used for representing content:

- TEI;
- JPEG;
- SGML;
- XML;
- DTDs.

The following standards, guidelines or tools are used for describing content:

- TEI header.

The program looked at other existing guidelines for digitizing particular document types when planning its digitization strategy. It rejected TEI as inadequate for editorial purposes.

The following standards, guidelines or tools are used for representing structure:

- TEI;
- HTML;
- JPEG;
- SGML;
- XML;
- several locally developed DTDs.

The DTDs were developed for the Archive by IATH. The principal of these is the Blake Archive Description (BAD) DTD. This is used to encode all works at the object and collection level. The TEI DTD is used for other materials in the Archive such as bibliographies and Erdman's *Complete Poetry and Prose of William Blake*.

The program does not make any recommendations in relation to standards in general, nor in terms of navigating between the ideal and the realistic.

The intended target audiences for digital deliverables are:

- K-12;
- community college;
- four-year college;
- graduate school;
- lifelong learning;
- distance learning;
- museum users;
- public library users;
- archives users.

The project carried out an evaluation of the target audiences. Various groups other than the primary target audience could used the digital deliverables. The profile of the users has been as anticipated.

35.2 Project Management and Planning

Experience in project management was available in-house. The program has led to changes in organizational relationships and procedures (no details available). There are formal project management procedures in place

(no details provided). No project management procedure has failed. Quality assurance procedures take the form of annual meetings to set the year's agenda, while grants require specific goals and program reports. These have been successful for grant applications.

Pilot studies have been carried out for training needs, technical feasibility, user needs, workflow analysis and technology forecasting. No significant changes have been made to the design of the project as a result. No time and motion or benchmarking studies have been carried out. Job descriptions are applied.

Digitization is carried out in-house using equipment that was already available and some that was bought in. The Archive uses a Microtek Scanmaker III and a Microtek Scanmaker V flatbed scanners (the latter uses a transparent media adapter) and a Nikon LS-3510AF slide scanner. Guidelines for data capture procedures have been established and color charts are the benchmarks used. The monitors and scanners are calibrated on a regular basis (using ProSense 1.8 software) and compressed air is used to remove dust and lint, from the scanner bed and transparency before every scan.

35.3 Human Resources and Training

The project employs three general editors (directors) plus a technical editor. One of the editors is also a digitizer; one or two graduate students are employed for approximately ten hours a week and a further two graduate students for markup also for ten hours a week. The project manager and technical support person are graduate students and work 15-20 hours per week. IATH has two technical development staff who work 5% of their time on the project. All the project staff have a humanities background and the markup students are specialized in Blake. A combination of external and in-house advice on the technical aspects of digitization is used (e.g. Worthy Martin, Kodak and Xerox for JPEG 2000).

Areas where training needs were identified were project management, application of technical standards, preparation and handling of materials for digitization, technical operation of digitization equipment, and metadata creation. The project director, technical staff and equipment operators engaged in training and this has been organized in-house (own consultants), through external courses, independent study and learning on the job. The training has met the needs of the project.

35.4 Project Life-Cycle Processes and Procedures

The project is aware of the copyright position of the digital deliverables and does not own the copyright in the original materials. Material in copyright was digitized with the owners' agreement (for which there are formal procedures). The program declares copyright or rights status. Users are allowed to make printouts of the digital deliverables on paper and download to a PC. Users can view texts marked up according to the Blake DTD and to TEI, but the actual markup is not available. Users can view lower and highest quality images. No electronic management systems, such as watermarking, are in use.

The project does not have a conservation procedure for the original materials (since this is not part of the project). The interviewee was uncertain whether the material's condition is investigated. The project does not modify, degrade or compromise material to carry out digitization. No information was available on risk assessment.

For illustrated books, the cataloging or reference system in place is the Bentley number system. There is a variety of catalogs and reference aids for non-illustrated books. Copy/edition information, location, etching print date, provenance from the catalog or reference system, and editor's information are used in the digitization process. The program does not alter originals for the digitization process or reject material. The Archive digitizes from three types of materials, two sizes of transparencies and slides. The intermediaries that have been used are 4x5 and 8x10 transparencies and occasionally 35mm slides. The project is developing an in-house system for non-illustrated books, possibly based on access accession numbers.

The Archive's second DTD is the Blake Object Description (BOD). This is used to encode the textual metadata comprising the Image Information Record. Each and every image in the Archive also contains textual metadata comprising its Image Information Record. The Image Information Record combines the technical data collected during the scanning process from the Image Production Record with additional bibliographic documentation of the image, as well as information pertaining to provenance, present location, and the contact information of the owning institution. These textual records are, at the most literal level, a part of the Archive's image files. Image

files are typically considered to be nothing but information about the images themselves (the composition of their pixilated bitmaps, essentially); but, in practice, an image file can be the container for several different kinds of information. The Blake Archive takes advantage of this by slotting its Image Information Records into that portion of the image file reserved for textual metadata. Because the textual content of the Image Information Record now becomes a part of the image file itself in such an intimate way, this has the great advantage of allowing the record to travel with the image, even if it is downloaded and detached from the Archive's infrastructure. Tools for controlling data values are controlled vocabulary and classification scheme.

35.5 Format, Resolution and Compression of Digitized Materials

The formats for retroconverted text-based digital deliverables are:

- SGML;
- XML.

Some texts contained non-Latin scripts. OCR is not used and keying in is the main conversion method for textual materials.

For image material the TIFF file format is used for capture and preservation and JPEG for delivery. The capture and preservation resolution is 600dpi. Delivery resolution is between 100 and 300dpi. Bit-depth is 24 throughout. Image compression is TIFF for preserving and JPEG for delivery. The aim of compression was to improve access, enhance usability and decrease storage requirements. The project retains the uncompressed scans. The project carries out post processing color correction on the images using PhotoShop and hooded Radius Press View 17SR and 21SR monitors. Color correction takes between 30 minutes and several hours. This is a key step to establishing the scholarly integrity of the Archive. Each transparency has first been color corrected to the original before the digital image is color corrected to the transparency. In this way the digital image will match the original when viewed under optimal conditions.

Average file size for capture and preservation is too variable to state. The quality control procedures in place for the digital deliverables are spot checks and total checks. Metadata quality control procedures are spot checks on the parsed files for images, while text files are checked by the editor.

Users can access the open catalog plus materials. Searching and browsing facilities are keyword and text description, Boolean, portions of the SGML, transcriptions and image markup. The Image Information Record may be viewed using either the "Info" button located on the control panel of the Archive's ImageSizer applet, or with the Text Display feature of standard software such as Adobe PhotoShop or X-View. Users do not have to pay for use of the digital deliverables.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, print media coverage, conferences and meetings, email shots and registering with search engines.

Special software used is Java and Dynaweb. Apart from searching and browsing, users are able to manipulate images by using an image sizer developed in house in Java. Usage levels are 1700 hits a day monitored by automatic data capture.

35.6 Evaluation, Funding and Long-term Sustainability

The program did not carry out front-end or formative evaluation of users before the development of the project. Summative evaluation was undertaken using email and observation of users' interaction with the system. The purpose of this evaluation was interface design. As a result, a revised interface was released in January 2001.

Funding sources were NEH, Getty, Mellon and sponsorship (Sun, IBM, INSO (for Dynaweb)). The program believes that the use of standards and guidelines in the digitization process has cost money. The program has been monitored by its funding through annual reports. Documents requested by funders include project management plans, cost models and workflow reports.

New material is added continually and the user interface updated regularly.

The Archive's preservation strategy is the use of durable and migratory encoding structures, and robust backup

and archive procedures. The Archive's project office retains hard copies of all Image Production Records, as well as ledgers tracking electronic file transfers, consignment of TIFF images to tape and CD-ROM, and shipping of transparencies and slides. Hard copies of the DTD and the Archive's SGML are also retained at the project office.

All of the Archive's data is safeguarded via daily incremental backups on magnetic tape. In the event of a catastrophic disk failure or a server break-in, the Archive's data could be quickly restored from the off-site backup system.

The digital deliverables will be available indefinitely and the project's longer-term sustainability is not reliant on self-generating funds.

Access tracking to all three of the Archive's servers is via Wusage 5.0.

The project does not have an exit strategy. Loss of the digital deliverable would be a matter of concern.

valid xhtml 1.0 strict
abp~04/02

36 Walker Art Center (WAC) and Minneapolis Institute of Art (MIA)

On October 20 2000, HATII interviewed Scott Sayre, Director of Media and Technology at the Interactive Media Group of the Minneapolis Institute of Art, and Robin Dowden, Director of Integrated Information Resources at the Walker Art Center. The Integrated Arts Information Access project had as its aim, the creation of an accessible interface for a standards-based information framework. The digitization process would lead to enhanced access to the permanent collections, archives and libraries of both the MIA and the WAC. It was intended that the resources could then be accessed for educational and other non-commercial purposes on the Internet. The interviews were conducted simultaneously and their responses have therefore been consolidated.

36.1 Organizational Digitization Program and Policy

The program and policy developed from the project on integrated art information access. Both the Walker and MIA distributed this through the web. They followed the AMICO course to digitize and structure information in line with standards where the standards existed. Both institutions have the same strategic goals. In MIA the digital media standards evolved quickly during the 1990s through the development of a series of multimedia projects.

In MIA the curatorial staff select approximately 125 items quarterly for digitization that have a scholarly or public interest.

The WAC has different criteria based on the smaller size of its collections — approximately 9000 objects as opposed to the 100,000 + objects in MIA. The visual arts creators identify an A-list of the top 25 or top 100. These are the “cream”, either for public or scholarly use. The photography and the digitization of the items are done in parallel, very often for a particular exhibition, for example, *100 years of sculpture*.

The priorities are established by the curators, rather than being driven by the digitization process in both institutions.

For both MIA and WAC the main obstacles are the movement of the art, either from storage, on loan or from the galleries. The scale of the work can also be a significant factor. Copyright is also recognized as an obstacle.

The main drivers for the selection of material for the WAC and MIA are, in order of priority: teaching and learning potential, enhanced access, improved functionality, social inclusion, potential to reach disadvantaged groups, and historical and cultural value. These are the main incentives and while other criteria may be a product of the digitization process, they do not influence the selection process. The WAC identified research significance as a further criterion of importance in its selection process.

These processes have not changed significantly over time.

The co-operation that took place in developing the digitization program was with museums and academic institutions. Most significant was the co-operation between WAC and MIA and with AMICO. The co-operation was local, regional, national and international mainly through the dialog with AMICO.

The WAC and MIA would encourage any institution to speak to other similar organizations before starting any digital program.

The current status of the project is ongoing with no anticipated end date. Both programs started in the early 1990s.

The main purposes in creating the digital deliverables for both WAC and MIA are to provide teaching and learning resource, public access, and wider access. For WAC, research and experiment are additional factors. MIA has identified cost savings as a long-term factor, where it is anticipated that the digital objects will be repeatedly usable for any purpose.

The program produced an explicitly informative statement of intent for the whole project, not specifically for the digital process. See http://www.walkerart.org/IAIA/ for all related documents.

The type of source material digitized includes:

- printed books, serials and similar bound materials;
- printed documents;
- typescript documents (MIA);
- unbound printed documents (MIA);
- printed illustrations;
- textiles;
- original artworks;
- intermediaries;
- photographic prints (MIA);
- sound broadcasts;
- video broadcasts;
- radio broadcasts;
- TV broadcast (WAC has done a little).

The last two are only ever done through intermediaries.

The nature and format of the materials digitized are extremely varied and include all sizes of photographs, original artworks and various sizes of documents.

Both MIA and WAC intend their digital objects to be used again in a variety of projects including online multimedia exhibitions.

The following standards, guidelines or tools are used for representing content:

- CIMI (WAC);
- JPEG;
- MPEG (MIA);
- SGML (WAC);
- XML;
- TIFF.

The following standards, guidelines or tools are used for describing content:

- MARC;
- Dublin Core;
- CDWA;
- VRA;
- AMICO.

The following standards, guidelines or tools are used for controlling data values:

- LC subject headings (not yet);
- AAT (considered)
- AMICO.

The following standards, guidelines or tools are used for representing structure:

- CIMI;
- HTML;
- XML (in the next two years).

Both MIA and WAC looked at standards and guidelines closely before starting the program. AMICO was the main tool used to develop the digital objects. Many of the standards available were not necessarily suitable for their needs, or else the time and expertise were not available to them.

The projects were aware of the possible guidelines and referenced them but developed their own for specific uses. They did not reject any guidelines as such. They advise an awareness of the larger community, which is often academic in nature, but also to be aware of the reality of the project and not to expect to find easy answers to problems.

Intended audiences for the digital deliverables include:

- K-12;
- community college;
- four-year college;
- graduate school;
- lifelong learning;
- computer mediated learning;
- museum users;
- public library users (by-product of digital process);
- archive users (by-product of digital process);
- government (by-product of digital process);
- public sector (by-product of digital process).

The WAC and MIA identified the K-12 group as their primary audience. The education sector as a whole is a high target with the other groups being more of an additional audience. They have produced material to support teachers' use of the online materials.

Evaluation reports based on usability labs have been produced. These can be found in the http://www.walkerart.org/IAIA/ site. They have no idea what the profile of the actual users is as this is an extremely difficult concept to measure.

There are no limitations to the use of the digital deliverables. However, they did not consult the W3C's Guide for accessibility for disabilities.

36.2 Project Management and Planning

Advice on managing the project was obtained from both in-house and external sources. MIA's digital program has had an impact on the photography department in the organizational structure, as it now has responsibility for the digitization process as well.

Both MIA and WAC have elaborate project management procedures including various committees with working groups, for example education, technical and project management. These procedures are in place through the IAIA program across both institutions. They have adjusted these procedures over time to be more inclusive.

The project management is part of the larger more formal organization. It is independent of any project and has layers of advisory groups, which report back to the institutions. This is not specific to the digital process.

The project carried out a sort of pilot study, a prototype that was used to identify user needs and technology forecasting. However, the whole program has been more of an organic process with each step taking into consideration the lessons learned from the previous steps. The prototype allowed the team to change the user interface, the display content, and the structure to accommodate the K-12 audience's needs. They did not undertake any benchmark studies.

There were no real Gantt charts or similar planning aids. The work is allocated based on the relationship to the existing job although some staff have changed jobs.

The digitization is carried out both in-house and is outsourced; the method depends on the materials. If material is fragile then the process will be carried out in-house. MIA outsources transparencies. This is to free up labor time and equipment. In-house methods enable further training of staff and control over the material and process.

Both institutions bought in equipment for the digitization process. The process is dictated by the nature of the material. Original artwork is photographed and the 4x5 slides are digitized. Flatbed scanners, film scanners and high-end digital cameras are used, a Phase One in particular. The film scanners tend to be Nikon Cool Scan and the flatbeds are either Agfa Duoscan or Lynz. Both MIA and WAC used gray scale and color charts, as well as Mac color sync profiles and Adobe gamma calibrator. Both have guidelines for the capture procedures.

36.3 Project Life-Cycle Processes and Procedures

Both MIA and WAC are aware of the copyright status of their materials. The institutions do not, on the whole,

own the original materials. They are owned by a variety of individuals or bodies, such as the artists or the artists' estates. They declare the copyright on the web page. WAC and MIA will generally get the owner's permission to use the original object, usually without formalities. WAC will digitize objects that are not in copyright and not use them, but to avoid disrupting the digital process they digitize until such times as the materials may be cleared for copyright.

The commercial usage of the material is strictly not permitted and there is a usage statement on the web page. Educational use is permitted.

In some cases, users can download RTF, PDF and Proprietary Word Processed documents; these tend to be project documents and not actual objects. The users can only view thumbnails and lower quality images. They can listen to 30-second samples and compressed sound as well as video samples and lower quality video clips. However, this will depend on the project.

They do not use watermarking as this is considered to be ineffective and costly and as yet there is no system that actually works well. There is no recognized standard format and using any system now would mean using a proprietary system.

Both institutions have a preservation or conservation procedure for original material, established by the conservators in the museums. They do not investigate the condition of the original materials nor do they undertake any conservation procedures specifically as part of the digitization process.

The risks identified to the material are handling and movement. Materials are not digitized if the process will cause significant risk. Intermediaries are used to reduce risk although this is less frequent with the purchase of the Phase One single shot camera. Cold lights are used for all photography, digital or traditional, with a variance in exposure for different materials. The conservation procedures used for photography are applied to the digital process.

The MIA curators will monitor and prepare the materials for the digital process. In WAC this will vary from department to department. Neither institution will restrict access to original materials but the long-term goal is to encourage users to access the digital surrogate and it is hoped that this will be an organic process.

An in-house catalog is shared with registration, archives and libraries. The program used the ID information, the accession number, the description and other information from the catalog. The MIA had the relevant materials before starting the digitization process while WAC enlarged the record as part of the process, locating some source materials to complete the record.

The WAC altered some objects before digitization as did the MIA, for example de-glazing artworks for photography. The tendency was not to reject material before digitization, although some objects were postponed until improvements in the technology decreased the risk of damage. Both institutions used intermediaries and original objects. The material has never only existed in reproduced form.

Both institutions reproduce their material mainly as 4x5 transparencies for images; audio is reproduced on digital CD or DAT and video copies from film. MIA also uses digital video and there is a growing number of born-digital objects.

The registration departments catalog both institutions' original objects. They have a different system for the digital surrogates based on VRA, MARC and EAD. They use Filemaker Pro software to build this catalog system. They have their own controlled vocabulary within their system. They record details about the digital object, the digitization process, technical details, staffing details (e.g. who scanned the object) and administrative details. The digitizer or photographer records these details.

The metadata are held in a separate catalog that is connected to the main catalog. It is available on the intranet server and available on the internet through AMICO. They are currently working on a new collections management system that will include digital objects in the main catalog.

Both institutions use XML, word processed files and TXT for retroconverted text. WAC uses HTML as well. They do not digitize non-Latin characters and do not use OCR. They use keying-in to convert their textual materials.

TIFF file format is used for capturing and preserving, as is Photo-CD in MIA. GIF, JPEG and Flash-Pix are used for delivery. The projects have standards on resolution and bit-depth.

Both MIA and WAC compress in JPG and FlashPix. Compression is for reducing cost, improving access, enhancing usability, decreasing storage and reducing theft of digital resources. It is the biggest method of risk

management for security of images and controlling the artists' license.

The project retains images in uncompressed form and carries out post processing color correction, cropping, and batch processing using PhotoShop. The average capture and preservation file size is 210MB and 10MB for delivery. The dynamic range of the equipment is checked using the built-in software of the equipment.

They would recommend capturing the greatest amount of information at scan time.

Both MIA and WAC digitize sound. MIA uses mainly QuickTime and WAC uses RealAudio. The details of the sampling rates and compression are forthcoming in a document to be sent to HATII.

Both MIA and WAC also digitize moving images. The details of these formats, sampling rates and compression are also forthcoming.

MIA and WAC have no formal quality control process but at image processing time each image is checked as part of this process, therefore 100% by hand. There is no real quality control process for metadata either, other than checking as part of the full process.

Access to the digital objects varies with the audience. External users have limited access to lower quality images. In-house users have access to the entire catalog and objects. Users can search across category, indexing for educational materials. Users can run unstructured queries and they can search on the associated metadata. Usage estimation is forthcoming. Users do not need any other software or hardware to access the materials. They can annotate, create sets and create detailed derivatives.

Users do not have to pay for the use of the digital deliverables.

Potential users of the digital deliverables are informed about their availability through website announcements, press releases, articles in print media, print and broadcast media coverage, conferences, meetings and electronic and conventional mail shots. The project believes that conferences and articles have been very successful in disseminating the information to the academic community but want to make efforts to have greater success with the K-12 audience.

36.4 Evaluation, Funding and Long-term Sustainability

No front-end evaluation has been done as such.

Formative evaluation took the form of questionnaires on paper and online, focus groups and observations of users interaction in a Usability Lab. This was video-recorded for later reference.

They changed the project from the ground up based on this evaluation. The structure, the descriptive vocabulary, display content and the user interface were all altered radically.

No summative evaluations have been undertaken as all activities are ongoing.

The projects realized that users' skills vary dramatically depending on the age and skill ratio. The evaluation allowed them to differentiate from their own ideas to those of users. Evaluation should be done as early as possible to allow feedback in the process, if possible true front-end evaluation. It should be remembered that users' needs and understandings can be difficult to gauge.

The funding for IAIA comes from the State department of education in a sizeable grant, approximately $2 million. While they have adequate funding it would be helpful to have a longer-term budget to ensure longevity and development. They agreed that using standards has saved money in the long-term as it allowed the collaboration between the WAC and MIA, sharing standards as well as ideas and staff.

There are quarterly reports on the IAIA website that are written for the funding organization.

MIA and WAC will continue to add new materials, update metadata and the user interface. They have plans to document formally the digitization process for the long term. They want to move into areas as yet untouched, such as the archives at the WAC.

There is no real preservation strategy other than as part of the larger organization's disaster recovery plan. WAC stores images on two CDs. They archive in TIFF and ASCII using CDs, and DATS.

They intend to keep the digital deliverables available for as long as possible and indeed the loss of this material would matter greatly.

36.5 Further Comments

It is the intention of the project to ensure that the digital process becomes part of institutional culture, to ensure that the resources are used day-to-day by all types of user. They wish to educate the staff to identify and use the objects in everyday work.

The WAC is bounded by the size of the collection, and wishes to move beyond the visual arts, which is only one track of the arts center.

From the press release on the project:

“The Minneapolis Institute of Arts (MIA) and the Walker Art Center have embarked on an ambitious joint project that eventually will bring their considerable visual and educational resources to classrooms and homes in Minnesota and the world via the Internet. A $1 million grant from the State of Minnesota Department of Children, Families & Learning launched the project in September 1997. Called the Integrated Arts Information Access project (IAIA), the multiyear effort will integrate the collections and archives of the two institutions in digital form.

The scope of the IAIA project will require an ongoing commitment from both the Walker and the MIA. During the initial phase, July 1, 1997, to June 30, 1999, both museums will concentrate on converting their collections and related resources to digital form. In addition to images, audio, video, and software, materials such as catalog records, curriculum, manuscripts, and print publications also will be converted. This work will pave the way for future development of online exhibitions, curriculum, and other activities aimed at local and remote K-12 students and families.”

http://www.walkerart.org/IAIA/12press.html

valid xhtml 1.0 strict
abp~04/02

37 WGBH Media Archives and Preservation Center & Digital Media Production

HATII interviewed Mary Ide, Director of the Media Archives and Preservation Center at WGBH, and Thom Shepard, Systems Programmer/Analyst, on January 25 2001. The Center is responsible for the construction and maintenance of the digital asset management system, which aims to preserve the considerable volume of material produced by WGBH, for the benefit of the public broadcasting industry. The advantages of digitally archiving the output of WGBH are largely industry-based, however, a subsidiary benefit is the potential for the material to be used for educational or research purposes.

37.1 Organizational Digitization Program and Policy

WGBH is public broadcasting's largest producer of nationally distributed programs and has been in the broadcasting business for 50 years. Their productions make up one-third of PBS prime-time lineup. Their radio productions are distributed nationally and internationally. The Media Archives and Preservation Center preserves WGBH television and radio programs, production media, photographic assets, data files and historical records. They hold approximately half a million items, including:

- approximately 30,000 cubic feet of archived materials (in formats ranging from posters and printed matter to computer discs, from digital audio tapes to a variety of video and film formats);
- more than 160,000 hours of television programming;
- more than 20,000 hours of film;
- more than 32,000 reels of fm radio programming.

A WGBH Media Archives and Preservation Center leaflet describes its organizational aims:

"Now, in the digital age of multimedia databases, online services and video on demand, WGBH has the responsibility – and the opportunity – to make this educational treasure available through new technologies, and to provide access to a growing inventory of original and archival footage, photographs, research materials and written transcripts. Preservation makes good business sense, too, allowing WGBH to meet the demands for rebroadcast, home-video sales and other emerging market opportunities for its programs."

Prioritization of digitization decisions in this area is quite different from the academic community. The primary driving forces are production and business opportunities and sales. Archival reasons follow. Institutional values, research reasons, re-use values, and cost of retention issues play a subsequent role. For example, institutional values can include:

- Does content relate to institutional values?
- Do we own it?
- Is there a significant amount of information about the content already available?
- Who was the source of funding?
- What were the reviews, the usage?
- Is this the only physical copy we have?

Re-use values might include:

- Can we make a potential sale of the film footage?
- Are there any multiversioning opportunities?

Cost of retention values might be:

- What is the cost of conversion?
- What is the cost of cataloging the item?
- What is the life expectancy of the current item?

One of the obstacles to planning and delivering the development of digital deliverables is funding (mainly internal in this case) – it is not easy to convince the organization to invest very large sums on archiving. This demands a high level of organizational cooperation and buy-on of the main idea. Staffing, technology and standards present continual challenges to be met.

Source material includes video, audio, film, paper, and computer discs.

In terms of data standards, the special projects use modified Library of Congress subject headings.

They are designing the digital asset management system primarily to support all operations of the organization. They can use it for internal research, re-broadcasting, selling videos and supporting educational programs.

37.2 Project Management and Planning

Twelve people work at Archives. Four of them are trained archivists, the rest come from a production background.

One of the quality assurance procedures they use is to review the data entry form.

37.3 Project Life-Cycle Processes and Procedures

All material is within copyright, and is generally a very complex area for this type of organization.

They are currently planning a pilot. They will carry out extensive analysis of the needs of the organization and analysis of the workflow.

There is a limited budget for reformatting material in-house. There are discussions at the moment about preservation.

The originals are cataloged using FileMaker and Oracle. The organization is currently trying to move everything to Oracle and integrate about half a million records in one system. This is a very messy process and they are trying to develop a standard vocabulary. This is not easy, as part of the data is often entered by busy people at the production floor, rather than archivists or information professionals. They are working on adapting Dublin Core.

They are currently testing Artesia Technologies' TEAMS digital asset management product. Current work includes modifying data structures and metadata fields to use in the new system. The first type of metadata is about Physical Control, where Elements include:

- tapes (originals, screeners, stock) and metadata fields are Tape ID, Tape Source, Label, BarCode, Format, Type;
- text (record group, record subgroup, record series, folder title, folder datespan, condition, etc);
- stills (picture number, picture type, format, filename, URL, Color Type, Quality, Photographer, Source, etc).

The second type concerns Intellectual Control, where *Programs* include information about:

- host department;
- series;
- episodes (cataloging information, people, summaries);
- segments.

37.4 Format, Resolution and Compression of Digitized Materials

From Betacam SP they convert to QuickTime, then compress for streaming with MediaCleaner and save on Digital Video Tape.

A pilot is currently underway, with documentaries for a web project with extra production material that was never included in the show. They are focus-testing this with targeted audiences.

Funding is internal by the organization.

37.5 Conclusion

WGBH is planning a major change in the way it handles its assets, both digital and analog, but the specifics at present are uncertain and it is undergoing a period of change and experimentation. In terms of digitization processes and management of metadata they are still in a developmental stage. There are many interesting issues that have to do both with the nature of the organization, which is very different from an academic or public cultural institution, and with the nature of the material.

valid xhtml 1.0 strict
abp~04/02

NATIONAL INITIATIVE FOR A NETWORKED CULTURAL HERITAGE

ninch

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

By the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and NINCH

Interview Reports
Interviews conducted by The Humanities Advanced Technology and Information Institute (HATII)

1 Introduction

2 The Bancroft Library

3 Brown University, Women Writers Project

4 The Online Archive of California

5 The Art Institute of Chicago, Imaging Department

6 University of Chicago, The Oriental Institute

7 University of Chicago Library

8 Colorado Digitization Project

9 Library of Congress, Geography and Map Division

10 Library of Congress, Preservation Reformatting Division

11 Library of Congress, Prints and Photographs Division

12 Library of Congress / National Digital Library Program (NDLP)

13 Cornell University

14 University of Illinois, Digital Imaging and Media Technology Initiative (DIMTI)

15 Images of England project (National Monuments Record, English Heritage)

16 Göttingen State and University Library, Center for Retrospective Digitization

17 Harvard University, Visual Information Access (VIA)

18 Harvard University Fine Art Museums / Collection Management System

19 Harvard University Library / Library Digital Initiative Program (LDI)

20 Indiana University, Cook Music Library

21 The Church of Jesus Christ of Latter-Day Saints, Genealogical Society of Utah (GSU)

22 University of Michigan

23 Museums and the Online Archive of California

24 New York Public Library (NYPL)

25 National Library of Norway

26 Survivors of the SHOAH Visual History Foundation

27 Indiana University, Thesaurus Musicarum Latinarum (TML)

28 Tufts University

29 Library of Virginia (LVA)

30 University of Virginia, Center for Digital History (Alderman Memorial Library)

31 University of Virginia, Electronic Text Center

32 University of Virginia, Institute for Advanced Technology in the Humanities (IATH)

33 University of Virginia, Robertson Media Center (Clemons Library)

34 University of Virginia, Special Collections (Alderman Memorial Library)

35 University of Virginia, William Blake Archive

36 Walker Art Center (WAC) and Minneapolis Institute of Art (MIA)

37 WGBH Media Archives and Preservation Center & Digital Media Production

38 Appendix -- The Interview Instrument

valid xhtml 1.1
abp~09/02

Digitization Data Collection Instrument for Site Visit Interviews

Preamble

Instructions for Interviewers

- fill in a name on the first sheet and give each interview a unique id in the form: 2512CD01 (interview conducted on 25th Dec, by CD, first interview)
- instructions in italics
- give interviewee a leaflet explaining the project and the aims and objectives of the survey
- give interviewee a Human Subject Consent Form

Definition of terms: ***Digital deliverable*** includes the software program, database, digital text, image, moving image or audio, multimedia or interactive application to be designed, created or written.

Sources
Surveys of this sort inevitably draw on previous similar studies. In addition to the Six Principles drawn up by the NINCH Working Group, we gratefully acknowledge the following sources:

1. Federal Task Force on Digitization, Towards a Learning Nation: The Digital Contribution, Recommendations Proposed by the Federal Task Force On Digitization, http://www.icsti.org/dtf-gtn/english/report/index.htm

2. IFLA Questionnaire on digitisation and preservation, http://www.ifla.org/ifla-qdp/eqdp.htm

3. InterPARES Project Authenticity Task Force Case Study Protocol Questionnaire http://www.interpares.org [although currently closed documentation].

4. Project Review Guidelines developed and used by HATII to assess projects for the Heritage Lottery Fund [currently closed documentation].

5. S.D. Lee, Scoping the Future of Oxford's Digital Collections, http://www.bodley.ox.ac.uk/scoping/

6. S. Ross, JISC Image Digitisation Management Models: A Review http://www.hatii.arts.gla.ac.uk/Projects/jidi

Site visit interview instrument

Name of project/program/organization/individual

……

Unique identification no **……**

[Day/month/initials/interview no.]

[Following sections to be answered by project manager/liaison]

Administration

Interview details

1. Date of interview:
2. Location of interview:
3. Interview conducted by:
4. Start time:
5. Finish time:

About the Organization & Project

6. Formal Name of Organization:
7. Name of Department (for visit):
8. Title of Project / Program:
9. Name Project / Program Director:
10. URL/other reference/further published info on project/program:

About the Person being interviewed

11. Name:
12. Job Title:
13. Address:
14. Telephone Number:
15. Fax Number:
16. Email:
17. URL:

Organizational digitization program and policy

[Not all of questions 18 to 30 may be applicable to individual projects.]

Collection Survey

18. How did your organization develop its digitization strategy? Could you describe the process or the activities that led to the development of the strategy?

19. Did your organization conduct a collection survey in the process of planning its digitization strategy?
Yes❒ No ❒ ***[If no, move to next section, question 25]***

If yes, ***[May we have a copy?]***

20. Was the institution as a whole made aware that the collection survey was being conducted and were all staff encouraged to participate?
Yes❒ No ❒

21. Who was involved in carrying out the collection survey?

22. What were the lessons learned from carrying out a collection survey?

23. If you were to conduct one now, would you do anything differently?

Priorities for digitizing

24. Was the collection survey used to establish priorities for digitizing holdings?
Yes❒ No ❒

If no, was there some other document that helped to establish priorities for digitization?

25. Who established the priorities for digitizing holdings?

26. Were these priorities formalized in a strategic policy statement?
Yes❒ No ❒

[If yes, may we have a copy?]

If no, do you have plans to develop a policy?
Yes❒ No ❒

If yes, to what timescale will this happen?
Timescale:

If yes, who prepared the formal strategic policy statement?

If yes, how was it approved within the organization?

27. What objectives did this policy attempt to achieve? ***(Please rank in order of importance where 1 is your most important objective)***

Objectives of Digitization Policy	Priority Ranking

28. Has the project been successful in achieving these objectives?

Yes❒ No ❒

If no, what were the obstacles?

29. What recommendations would you have for other organizations that are attempting to formalize their selection criteria in strategic policy statements?

Overall obstacles

30. Can you identify the obstacles to planning the development of digital deliverables from your collections?

31. Can you identify the obstacles to the process of building digital deliverables from your collections?

Selection and Prioritization of materials

32. Which of these criteria guided you in the selection and prioritization of materials for digitization? ***(Please tick all that apply and rank in order of importance where 1 is your most important criterion***)

Selection Criteria for Digitization	Tick	Selection Rank	Prioritization Rank
Intellectual Property Rights			
Teaching & Learning potential			
Conservation (e.g. decreased wear and tear)			
Historical and cultural value			
Potential commercial exploitation			
Research significance			
Provision of user services (e.g. document delivery)			
Improved functionality			
Enhanced access			
Research into digitization strategies			
Publicity			
Space rationalization			
Preservation (e.g. against content loss)			
Labor cost reduction (e.g. reducing curatorial intervention)			
Potential to reach disadvantaged groups			
Infrastructure cost reductions			
Social Inclusion			
Other (please specify)			

33. Did you find that your selection criteria changed over time?
Yes❒ No ❒

If yes, what were the changes that occurred?

34. Have your prioritization criteria changed over time?
Yes❒ No ❒

Co-operation

35. Did you co-operate with other organizations to develop your digitization program? ***(Please tick all that apply)***

Type of Organizations you Co-operated with for Digitization	Tick
Archives	

Libraries	
Museums	
Academic Institutions	
Corporations	
Foundations & Charities	
Government Agency	

36. Was this co-operation?

Type of Co-Operation	Tick
Local	
Regional	
National	
International	

37. Based on this experience, what recommendations would you have for institutions thinking of starting collaborative digitization projects or programs?

Project- or Program-Related Questions

38. What is the current status of the project/program?
Ongoing ❒ Completed ❒

39. Please give:

Start date:
End date (if completed):
Anticipated end date (if not yet completed):

Purpose of the project / program

40. For what purpose were the digital deliverables created? ***[If more than one purpose, please rank in order of priority where 1 is your most important purpose.]***

Purpose of Digitization Project / Program	Rank
Preservation	
Public access	
Teaching and learning resource	
Research	
Experiment	
Wider access	
Revenue generation	
Response to previous demand	
Other (give details)	

41. Did the project/program produce an explicit informative statement of intent (e.g. a publicity document, an announcement, on a website)?

Yes❒ No ❒

If yes, was this statement explicit about its: ***(Please tick all that apply)***

Content of Statement of Intent for Digitizaton Project / Program	Tick
Rationale	
Scope	
Significance	
Primary audience	
Long-term sustainability	

Level of faithfulness to an original and/or intermediate, whether analogue or digital	
Suitability for different target audiences	

[May we have a copy or reference to it?]

Nature of the source materials and the impact of digital deliverables

42. What type of source material was digitized? ***[Please tick all that apply]***

Type of Source Material Digitized	Tick
Printed books, serials, and similar bound materials (published)	
Printed documents (grey literature)	
Unbound printed documents	
Typescript documents	
Printed illustrations	
Handwritten documents	
Artifacts	
Original artworks	
Intermediaries (film, microfilm, microfiche, photocopies etc.)	
Mechanical copies	
Hand copies	
Photographic Negatives (but not intermediaries)	
Photographic Prints (but not intermediaries)	
Textiles	
Film	
Sound recordings	
Video recordings	
Radio Broadcasts	
TV Broadcasts	
Other ***(please specify)***	

43. What was the nature and format of the materials to be digitized? ***[e.g. medium and size]***

44. Were the digital deliverables intended to represent a sample of a larger body of material, or were they a representation of an entire body of material?

Sample ❒ Entire body of material ❒

45. Were the digital deliverables intended to be re-used/re-purposed as part of the project/programme (e.g. in an educational authored CD-ROM)?
Yes❒ No ❒

If yes, please give details.

46. Did this approach work well?

47. What would you change if you were to do it again?

Interoperability

48. Did you use any of the following standards, guidelines, or tools for **representing content**?
Yes❒ No ❒

If yes, please identify the standards used ***(tick all that apply)***

Standards, Guidelines, or Tools for Representing Content	Tick
TEI	

CIMI (old CIMI DTD or Dublin Core DTD) ***please specify***	
Unicode	
JPEG	
MPEG	
SGML	
XML	
HyTime	
SMDL	
Laban notation	
Other ***(please specify)***	

If no, can you explain why not? ***(Please tick all that apply)***

Reasons for Not Using Standards for Representing Content	Tick
Standards not yet agreed	
Standards not relevant to type of material	
Standards not known	
Lack of documentation about standards	
Lack of training about standards	
Not appropriate to the project	
Other ***(please specify)***	

49. Did you use any of the following standards, guidelines, or tools for **describing content**?
Yes❒ No ❒

If yes, please identify the standards used ***(tick all that apply)***

Standards, Guidelines, or Tools for Describing Content	Tick
MARC	
Dublin Core	
EAD	
TEI Header	
Categories for the description of Works of Art	
Other ***(please specify)***	

If no, can you explain why not? ***(Please tick all that apply)***

Reasons for Not Using Standards for Describing Content	Tick
Standards not yet agreed	
Standards not relevant to type of material	
Standards not known	
Lack of documentation about standards	
Lack of training about standards	
Not appropriate to the project	
Other ***(please specify)***	

50. Did you use any of the following standards, guidelines, or tools for **controlling data values**?
Yes❒ No ❒

If yes, please identify the standards used ***(tick all that apply)***

Standards, Guidelines, or Tools for Controlling Data Values	Tick
AAT	
Library of Congress subject headings	
Library of Congress subject thesaurus	
Name authority files	
Other ***(please specify)***	

If no, can you explain why not? ***(Please tick all that apply)***

Reasons for Not Using Standards for Controlling Data Values	Tick
Standards not yet agreed	
Standards not relevant to type of material	
Standards not known	
Lack of documentation about standards	
Lack of training about standards	
Not appropriate to the project	
Other ***(please specify)***	

51. Did the project/program look to existing guidelines to digitizing particular document types and object types when planning its digitization strategy?

Yes❒ No ❒ ***[If no, go to question 53]***

If yes, which guidelines did you adopt?(e.g. TEI/MASTERS proposals for manuscripts; LC/Ameritech Competition Guidelines)?

If yes, what modifications did you carry out to make these suitable to your project/program?

52. Which guidelines did you reject, and why?

53. Did you use any of the following standards, guidelines, or tools for **representing structure**?
Yes❒ No ❒

If yes, please identify the standards used ***(tick all that apply)***

Standards, Guidelines, or Tools for Representing Structure	Tick
TEI	
CIMI (old CIMI DTD or Dublin Core DTD) please specify	
HTML	
Unicode	
JPEG	
MPEG	
SGML	
XML	
HyTime	
SMDL	
Laban notation	
Other ***(please specify)***	

If no, can you explain why not? ***(Please tick all that apply)***

Reasons for Not Using Standards for Representing Structure	Tick
Standards not yet agreed	
Standards not relevant to type of material	
Standards not known	
Lack of documentation about standards	
Lack of training about standards	
Not appropriate to the project	
Other ***(please specify)***	

54. In relation to the use of standards in general, would you have any recommendations for navigating between the ideal and the realistic or the necessary?

Target audience

55. What was the intended audience for the digital deliverables? ***[Please tick all that apply and rank in order of priority***]

Intended Audience for Digital Deliverables	Tick	Priority Ranking
Teaching and learning:		
K-12		
Community college		
Four-year college		
Graduate School		
Lifelong learning		
Distance learning		
Computer-mediated learning		
Other (please specify)		
General public:		
Museum users		
Public library users		
Other ***(please specify)***		
Subject specialists:		
Archive users		
Other ***(please specify)***		
Other:		
Government		
Private sector		
Other ***(please specify)***		

56. Did you conduct any evaluation of the target audience and its needs?
Yes❒ No ❒
[If yes, could we have a copy of the relevant report, published article, etc?]

57. Could the digital deliverables be used by any users other than the primary intended target audience?
Yes❒ No ❒

If yes, which other target audiences might use the digital deliverables?

58. Did the project acknowledge the needs of those with disabilities or take account of W3C's "Guidelines for Web Site Accessibility" (http://www.w3.org/WAI/)?
Yes❒ No ❒

If yes, in what way?

59. If limitations as to use of the digital deliverables were necessary, did you clearly state the reasons for this?
Yes❒ No ❒

If yes, how and where did you state the limitations

60. What was/were the reason(s) for these limitations?

61. Was the profile of actual users the same as the one anticipated?
Yes❒ No ❒

Project Management and Planning

Project Management

62. Did/do you use any external advice and consultancy on managing the digitization project/program, or was/is advice available in-house?
External ❒ In-house ❒

63. How is the management of the digitization project/program related to the structure of the larger organization?

64. Did the digitization project/program lead to any changes or re-evaluation of organizational relationships and procedures?
Yes❒ No ❒

65. Did you have formal project management procedures in place? (e.g. Project Management Board/Steering Group?)
Yes❒ No ❒

If yes, please specify

66. Have you used any project management procedures that did not work?
Yes❒ No ❒

If yes, can you explain why?

67. What managerial Quality Assurance procedures have you put in place?

68. What steps did you follow to develop these Quality Assurance procedures?

69. Were these successful?
Yes❒ No ❒

Can you explain why or why not?

Project Planning

70. Did you carry out a feasibility or pilot study to provide information to assist project management and planning?
Feasibility ❒ Pilot Study ❒ Both ❒ Neither ❒

If yes, which of these was it used for? ***(Please tick all that apply)***

Purpose of Feasibility or Pilot Study	Tick
Scheduling	
Training needs	
Technical feasibility	
User needs	
Workflow analysis	
Workflow piloting	
Technology forecasting	
Other ***(please specify)***	

[If there is any summary report of the pilot/feasibility study, could we have a copy?]

71. Did you make significant changes to the design of the project/program as a result of any preliminary study?
Yes❒ No ❒

If yes, in what areas were these changes made?

72. Did you undertake any time and motion or other benchmarking study at the start of the project/program to determine the viability of scheduling tasks?

Yes❒ No ❒

If yes, what parameters were taken into account in time and motion planning and benchmarking?

What types of tools did you use to aid planning and scheduling (for example, did you use organizational charts or Gantt charts)?

73. How did you determine who would do what work?

74. Did you produce job descriptions and performance indicators for each

Job description ❒ Performance indicators ❒ Both ❒ Neither ❒

Planning digitization

75. How was digitization carried out?
In-house ❒ Outsourced ❒

Could you please give reasons for your decision?

76. If digitization was done in-house, did you already have access to digitization equipment or did you buy in equipment for the project/program?
Equipment already installed ❒ Equipment bought in ❒

77. How did you choose which digitization process to adopt?

78. Did you use any of the following technologies? ***(Please tick all that apply)***

Technologies Used for Image Digitization	Tick	Specify Manufacturer
Flatbed scanners		
Drum scanners		
Film scanners		
Digital cameras		
High-end professional cameras (e.g. Kontron, Phase One)		
Other (please specify)		

79. Did you establish guidelines for data capture procedures (e.g. calibration of equipment, handling dust etc.)?
Yes❒ No ❒

80. Which benchmarks did you select? ***(Please tick all that apply)***

Benchmarks Selected for Image Digitization	Tick
Gray scales	
Color charts	
EDG charts	
Reproduction charts	
Other (please specify)	

Human resources and training

81. How many people worked on the project/program ***(in full-time equivalents)*** and in what capacities?

Type of Staff	Number	% of time on

		the project
Director		
Metadata specialist		
Curator		
Digitizer		
Photographer		
Technical support staff		
Technical development staff		
Education specialist		
Evaluation specialist		
Other (please specify)		

82. What kind of background did they have?

83. Were the staff re-deployed from other areas?
Yes❒ No ❒

84. Did you use any external advice and consultancy on technical aspects of digitization, or was advice available in-house?
External ❒ In-house ❒

85. How did you assess the training needs of the project team?

86. In which areas were training needs identified? ***(Please tick all that apply)***

Areas where Training is Needed	Tick
Project management	
Application of technical standards	
Preparation and handling of materials for digitization	
Technical operation of digitizing equipment	
Post digitization processes, e.g. editing	
Metadata creation	
Digital preservation	
Other (please specify)	

87. Which team members received training? ***(Please tick all that apply)***

Team Members Engaged in Training	Tick
Project/program Director	
Curatorial staff	
Specialist technical staff	
Conservation Staff	
Library/cataloguing staff	
Equipment Operators	
Other (please specify)	

88. How was training organized? ***(Please tick all that apply)***

Type of Training Organized	Tick
In-house, using project staff	
In-house, using organization's own consultants	
In-house, using external consultants	
Attendance at external courses	
Independent study	
Learning on the job	
Other (please specify)	

89. Did the training meet the needs of the project/organization?
Yes❒ No ❒

Can you explain why or why not?

Project Life-Cycle Processes and Procedures

Reproduction and copyright

90. Are your aware of the copyright position or other rights status of the digital deliverables that your project/program created?
Yes❒ No ❒

91. Does your organization own the copyright in the original materials?
Yes❒ No ❒

If no, who does own the copyright?

92. Did you declare the copyright or other rights status of the final digital deliverables?
Yes❒ No ❒

If yes, how did you do so?

Was this practice effective?
Yes❒ No ❒

93. Did you digitize material in copyright?

If yes, was this done: ***(Please tick all that apply)***

Way of Organizing Digitization of Material in Copyright	Tick
Under legal provisions for libraries	
With the owner's agreement	
By payment to the owner of a fee	
Under license	
Without formalities	

94. Are users of the digital deliverables allowed to do any of the following? ***(Please tick all that apply***)

Copying Facilities For Users	Tick
Make printouts on paper	
Print on film	
(for audio) Burn to CD, DVD, etc	
Download to a PC	
Download to a local network (LAN)	
Download to a general network (WAN)	

95. What variety of digital deliverables do you allow users to download? ***(Please tick all that apply)***

Type of e-text	**Downloading**	**Viewing only**
ASCII Text Files		
TEI DTD Marked-up text		
XML Marked-up		
RTF		
PDF		
Proprietary Wordprocessed		

Type of Digital Image Available for	Downloading	Viewing only
Thumbnails		
Lower quality images		
Highest quality images		
Associated documents		

Type of Digital Audio Available	Downloading	Listening only
Less than 30 second Samples		
Full length compressed		
High fidelity sound		
Associated documents		

Type of Digital Moving Image Available	Downloading	Viewing only
Samples		
Lower quality digital video clips		
Highest quality digital video clips		
Associated documents		

96. Were any electronic management systems used to control copying, e.g. digital watermarking?
Yes❒ No ❒

If yes, which management systems?

97. Were these electronic management systems effective?
Yes❒ No ❒

If no, why not?

Preservation/Conservation

98. Did your organization have a conservation procedure or policy for original materials?
Yes❒ No ❒

99. What investigations into the condition of the original materials did you do prior to digitization?

100. What conservation activities were undertaken on original materials as part of the project/program?

101. What risk to the material (if any) was identified during the preparation or digitization process?

102. How was the risk assessed?

103. What steps did you take to minimize the risk?

104. Did you specify special equipment and processes for vulnerable materials (such as special cradles, lighting and exposure limits)?
Yes❒ No ❒

If yes what special equipment or processes did you use?

105. Were the materials prepared by curatorial or preservation staff before digitization?

Yes❒ No ❒

106. Were the materials prepared/monitored by curatorial or preservation staff during digitization?
Yes❒ No ❒

107. Once the material has been digitized, will (did) the project/organization restrict access to the original?
Yes❒ No ❒

If yes, what criteria were/will be used to allow or restrict access to original materials?

Preparation & Sources

108. What cataloguing or reference systems were in place before digitization?

109. What information was used from these systems in the digitization process?

110. Did you have access to all the relevant materials (e.g. cataloguing information) for your digital deliverables collected before the digitization process?
Yes❒ No ❒

111. Did you have to locate any core reference or source materials to complete the digital deliverables?
Yes❒ No ❒

112. Did any material have to be altered from the original form for the digitization process, e.g. disbinding, removal from frames?
Yes❒ No ❒

113. Did you reject material before digital imaging?
Yes❒ No ❒

If yes, can you give reasons why? ***(Please tick all that apply)***

Reasons for Rejecting Material for Digitization	Tick
Skewing	
Incoherent pagination	
Incomplete works	
Print skew	
Margins narrower than 1 cm	
Major variation of density in same micro-image	
Micro-images of inappropriate density	
Blurred pages	
High percentage of illustrations	
Other (please specify)	

114. Did the project/program digitize originals or did it use reproductions/ intermediaries?
Originals ❒ Reproductions/ Intermediaries ❒

115. Did the material only exist in reproduced form?
Yes ❒ No ❒

116. What form(s) of reproduction/intermediaries did you use? (***Please tick all that apply)***

Form of Reproduction/ Intermediaries Used for Images	Tick

Photocopies	
Slides, 35 mm or 4 * 5 transparencies	
Photographic prints (e.g. photograph)	
Microfilm	
Single frame microfiche	
Others (please specify)	

Form of Reproduction/ Intermediaries Used for Audio	Tick
78 rpm discs	
33 rpm discs	
Audio cassettes	
Wax cylinders	
Digital sound media (CD, DAT)	
Others (please specify)	

Form of Reproduction/ Intermediaries Used for Moving Images	Tick
Video copies from film	
Betacam SP	
Others (please specify)	

Metadata

117. Do you catalog the original (physical) materials, if so how?

118. How do you catalog digital surrogates? (same system or different?)

Did you use any of the following standards or guidelines for cataloging the digital deliverables? ***(Please tick all that apply)***

Standards or Guidelines Used for Cataloging the Digital Deliverables	Tick
Dublin Core	
MARC	
RDF	
EAD	
TEI Header	
USMARC	
VRA	
Other (please specify)	

If no, did you create your own in-house system?
Yes ❒ No ❒

Why, and what software/system did you use?

119. Which tool did you use for controlling data values (e.g. controlled vocabulary, thesaurus, classification scheme)? ***(Please tick all that apply and specify the tool))***

Tool Used for Controlling Data Values	Tick	Please Specify
Controlled vocabulary		
Thesaurus		
Classification scheme		
Other		

120. Did you record details about: ***(Please tick all that apply)***

Metadata Recorded	Tick
The original object	

The digital object	
The digitization process	
Technical details	
Staffing details	
Administrative information	

121. Who created the metadata records? ***(Please tick all that apply)***

Specialization of Metadata Creator	Tick
Digitizer	
Archivist/information professional	
Digitizer who was also Archivist/information professional	
Other (please specify)	

122. Were the metadata records for the digital deliverables included in the main catalog or in a separate catalog?
Metadata included in main catalog ❒ Metadata in separate catalog ❒

123. Was the catalog for the digital deliverables: ***(Please tick all that apply)***

Form of Catalog for Digital Deliverables	Tick
In paper form	
In electronic form on an intranet server	
Available on the Internet (e.g. websites)	
Other (please specify)	

124. Were the records for digital deliverables and the original digitized materials: ***(Please tick all that apply)***

Relationship of Original Digital Deliverables and Associated Records	Tick
The same	
Independent of each other	
A mixture of the above	
Other (please specify)	

125. How are the catalog and the objects linked?

Format, resolution and compression of digitized materials

A. Text

126. What was the format chosen for retroconverted text-based digital deliverables? ***(Please tick all that apply and specify the details)***

Format of Text-based Digital Deliverables	**Tick**	**Please specify type, level, approach to markup**
HTML markup		
SGML markup		
XML markup		
PDF		
txt (ASCII)		
Word Processed		
Database		
Spreadsheet		
Postscript		
EPS (Encapsulated Postscript)		
Other (***please specify)***		

127. Did the documents digitized contain characters in non-Latin scripts?
Yes ❒ No ❒

128. Did you use OCR software to convert the digital images?
Yes ❒ No ❒

If yes, which software was used?

If yes, what level of accuracy did you achieve?

[If no, go to question 133]

129. Was any special treatment carried out prior to OCRing?
Yes ❒ No ❒

130. What was the aim of OCRing the material? ***(Please tick all that apply)***

Aim of OCR	Tick	Please describe
Automatic Indexing		
Enhanced searching		
Computer-based analysis		
Other (please specify)		

131. From your experience, what advice would you give to another organization that was about to start OCRing documents?

132. Did you use "keying in" as the conversion method for any of your textual materials?
Yes ❒ No ❒

If yes, what were the lessons learned from this experience?

B. Images

133. What file format did you use for **capturing, preserving, and delivering** the images? ***(Please tick all that apply)***

Image File Format	Capturing	Preserving	Delivering
GIF			
TIFF			
JPEG			
JPEG 2000			
Flash-Pix			
Photo-CD			
PNG			
Postscript			
PDF			
XBM			
Other (Please Specify)			

134. What resolution (in dpi) did you use for capturing, preserving, and delivering the images?

Image Resolution	Capturing	Preserving	Delivering
Resolution in dpi			

135. What bit depth did you use for capturing, preserving, and delivering the images? the project/program adopt?

Image Bit Depth	Capturing	Preserving	Delivering
Bit depth			

136. Did you use any compression method at capturing, preserving, and delivery stage?
Yes ❒ No ❒ ***[If no, go to question 139]***

If yes, what was the compression method used?

Image Compression	Capturing	Preserving	Delivering
Compression (please specify)			

137. What was the aim of the compression? ***(Please tick all that apply)***

Aim of Compression	Tick
To reduce cost	
To improve access	
To enhance usability	
To decrease storage requirements	
To reduce the risk of theft of digital resources	
Other (please specify)	

138. Does the project/program retain the original scans in **uncompressed** form?
Yes ❒ No ❒

139. Did the project/program carry out processing on the images?
Yes ❒ No ❒

If yes, please specify the type of processing and software used?

140. What were the average file sizes created at capturing, preserving, and delivery stages?

Image File Sizes	Capturing	Preserving	Delivering
Average File Size			

141. Did you check the dynamic range of your scanning equipment?
Yes ❒ No ❒

If yes, how did you do so?

142. From your experience with digitizing images, what recommendations would you make to colleagues starting work in this area?

C. Sound

143. Did the project/program digitize sound materials?
Yes ❒ No ❒ ***[If "no", go to question 154, next section]***

144. What work methods did you use for the digitization of sound?

145. What file format did you use for capturing, preserving, and delivering sound-based digital materials? ***(Please tick all that apply)***

Audio File Format	Capturing	Preserving	Delivering
WAV			
AIF			
RealAudio			
MID			

Ra			
Au			
MPEG			
QuickTime			
Other (please specify)			

146. What sampling rate did you use for capturing, preserving, and delivering sound-based digital materials (e.g. 11.025 kHz, 22.05 kHz, 44.1 kHz, or 48 kHz)?

Audio Sampling Rate	Capturing	Preserving	Delivering
Sampling Rate			

147. What bit-rate did the project/program use for capturing, preserving, and delivering sound-based digital materials (e.g. 8-bit, 16-bit mono, 16-bit stereo, 20-bit stereo)?

Audio Sampling Rate	Capturing	Preserving	Delivering
Bit Rate			

148. Did you use any compression method at capturing, preserving, and delivery stage?
Yes ❒ No ❒ ***[If no, go to question 152]***

If yes, what was the compression method used?

Audio Compression Method	Capturing	Preserving	Delivering
Bit Rate Compression (please specify)			
Frequency Compression (please specify)			

149. What was the aim of the compression? ***(Please tick all that apply)***

Aim of Compression	Tick
To reduce cost	
To improve access	
To enhance usability	
To decrease storage requirements	
To reduce the risk of theft of digital resources	
Other (please specify)	

150. What were the average file sizes created at capturing, preserving, and delivery stage?

Audio File Size	Capturing	Preserving	Delivering
Average File Size			

151. Was post-capture editing carried out?
Yes ❒ No ❒

If "yes", what type?

D. Moving image

152. Did the project/program digitize moving images?
Yes ❒ No ❒

[If "no", go to question 160 next section]

153. What work methods did you use for the digitization of moving images?

154. What file format did you use for capturing, preserving, and delivering moving image-based digital deliverables? ***(Please tick all that apply)***

Moving Image File Format	Capture	Preservation	Delivery
AVI			
MPEG (which version?)			
QuickTime			
RealVideo			
MOV			
Other (please specify			

155. What was the resolution that the project/program used?

Moving Image Resolution	Capturing	Preserving	Delivering
Resolution (frames per second)			

156. Did you use any compression method at capturing, preserving, and delivery stage?
Yes ❒ No ❒ ***[If no, go to question 160]***

If yes, what was the compression method used?

Moving Image Compression	Capturing	Preserving	Delivering
Compression (please specify)			

157. What was the aim of the compression? ***(Please tick all that apply)***

Aim of Compression	Tick
To reduce cost	
To improve access	
To enhance usability	
To decrease storage requirements	
To reduce the risk of theft of digital resources	
Other (please specify)	

158. What were the average file sizes created at capturing, preserving, and delivery stage?

Moving Image File Size	Capturing	Preserving	Delivering
Average File Size			

159. Did the project/program carry out any post-capture processing on the moving images?

Yes ❒ No ❒

If yes, please specify the type of processing and software used?

Quality Control

160. Do/Did you have any procedures in place for ensuring the quality of the digital deliverables?
Yes❒ No ❒

If yes, what kind of procedures? ***(Please tick all that apply)***

Type of Quality Assurance Procedures	Tick
Spot checks	
Random set of checks on materials	
Check on a stratified random sample	

Percentage checks on carriers (e.g. corrupted disks etc.)	
Total Check	
Other (please specify)	

161. Do/Did you have any procedures for ensuring quality control for metadata recording and delivery?
Yes❒ No ❒

If yes, what kind of procedures?

162. How do/did Quality Control procedures affect your workflow and policy decisions?

Delivery, access and use of digital deliverables

163. How do users access the digital deliverables? ***(Please tick all that apply)***

Type of User Access to Digital Deliverables	Tick
Open access to catalogue	
Open access to catalogue plus materials	
Access restricted to in-house users	
Access restricted	
Other (please specify)	

164. Which search/browsing facilities does/did the system you used offer?

165. How does/did your system handle searching/browsing of associated metadata?

166. Does/Did your digitization project/program have any special hardware/software needs? (e.g. access control mechanism, billing software)
Yes ❒ No ❒
If yes, please specify

167. Apart from browsing the digital deliverables, does/did your system offer users the opportunity to manipulate them in any way (e.g. save or annotate found sets, run OCR software on documents, edit sounds or moving images)?
Yes ❒ No ❒
If yes, please specify

168. What is/was the level of use of the digital deliverables?

169. Is/Was usage monitored via

Automatic data capture ❒ Formal data collection ❒ Informal observation ❒

Usage Charges

170. Do users have to pay to use the digital deliverables?
Yes ❒ No ❒

If yes, which types of users have to pay? ***(Please tick all that apply)***

Type of Users Charged for Use of Digital Deliverables	Tick
On-site users	
Users outside the organization	
General Public	
Other (please specify)	

171. If charges are/were made, how are/were these calculated? ***(Please tick all that apply)***

Way of Calculating Charges for Digital Deliverables	Tick
Single charge	
By time	
By volume of material	
By intended use (commercial/academic/students)	
Other (please specify)	

172. If charges are/were made how are/were they collected? ***(Please tick all that apply)***

Way of Collecting Charges for Digital Deliverables	Tick
Invoice	
Cash at point of use	
Credit card by telephone	
E-commerce	
Other (please specify)	

173. Do/Did you estimate the revenue generated by the use of the digital deliverables before commencing the project?

Yes ❒ No ❒

If yes, what is the level of revenue?

174. Did the projected revenues accurately anticipate the revenues generated through charging?

Yes ❒ No ❒

Dissemination and Publicity

175. How are/were potential users informed about the digital deliverables that you create(d)? ***(Please tick all that apply)***

Method of Publicizing Digital Deliverables	Tick
Announcement on website	
Press release	
Articles in print media	
Advertising in print media	
Print and broadcast media coverage	
Conferences, meetings	
Electronic mail shot	
Conventional mail shot	
Registering with web search engine(s)	
Other (please specify)	

176. Which of the dissemination and publicity strategies that you followed proved to be most effective?

Evaluation

Front-end / formative / summative evaluation of project/program and of product

177. Did you carry out any front-end evaluation with users before the development of the project/program?

Yesr No r ***[If no, move to question 179]***

If yes, which techniques did you use for gathering front-end evaluation data? ***(Please tick all that apply)***

Front-End Evaluation Techniques Used	Tick
Questionnaires on paper	
Questionnaires online	
Email	
Focus group discussions	
Observation of users' interaction with the system	
Computer interaction logging (of users' choices and navigation)	
Other (please specify)	

178. Did you change any aspect of the project/program as a result of the front-end evaluation findings?
Yes r No r

If yes, what did you change?

179. Did you carry out any formative evaluation with users during the design of the project/program?
Yesr No r ***[If no, move to question 181]***

If yes, which techniques did you use for gathering formative evaluation data? (***Please tick all that apply)***

Formative Evaluation Techniques Used	Tick
Questionnaires on paper	
Questionnaires online	
Email	
Focus group discussions	
Observation of users' interaction with the system	
Computer interaction logging (of users' choices and navigation)	
Other (please specify)	

180. Did you change any aspect of the project/program as a result of formative evaluation?
Yes r No r

If yes, what did you change?

181. Did you carry out any summative evaluation with the users after the project/program had been completed?
Yes r No r ***[If no, move to question 186]***

If yes, which techniques did you use for gathering summative evaluation data?
(Please tick all that apply)

Summative Evaluation Techniques Used	Tick
Questionnaires on paper	
Questionnaires online	
Email	
Focus group discussions	
Observation of users' interaction with the system	
Computer interaction logging (of users' choices and navigation)	
Other (please specify)	

182. What was the purpose of summative evaluation?

183. Is there any aspect of the evaluation work that you would carry out differently if you were doing it again?

184. What were the lessons learned from the evaluation work?

185. Did you make available or disseminate the result of these studies?
Yes r No r

If yes, how?

[May we have a copy of any evaluation report, publication, etc?]

Funding

186. How much did this project/program cost so far and total?

187. What were the funding sources for the project/program?

188. How much funding would you like to have/had?

189. If you had that much funding, what would you do differently?

190. If you used standards, do you feel they saved you money, or cost you more to implement, both short and long-term?

191. How did the funding organization monitor your project? (e.g. where any impact measures used to evaluate the results?)

192. Did the funding organization ask you to provide any of the following?
(Please tick all that apply)

Documents Required by Funding Organization	Tick
Collections Survey	
Conservation Plan	
Risk Assessment	
Front-end Evaluation report	
Project Management Plan	
Cost Models	
Workflow Reports	
Other (please specify)	

Long-term sustainability

193. Which of the following elements of the project will need to be updated and how often?
(Please tick all that apply and specify frequency)

Project Element to be Updated	Tick	Frequency
New materials digitized		
New metadata added		
Metadata changed		
User interface changed		
File Formats changed		
Other (please specify)		

194. Has the project/program or organization have a preservation strategy to ensure long-term access to documents and objects in digital form?
Yes r No r

If yes, what are the components of the preservation/archiving strategy? (e.g. in relation to workflow, staffing, resource issues)

If there are different strategies for different types of content, please describe the relevant strategy for each:

Preservation/Archiving Strategy	Text	images	Audio	Moving Images
File Formats				
Storage Media				
Storage conditions				
Updating Metadata				
Other (please specify)				

195. Are these preservation strategies based on: ***(Please tick all that apply)***

Approach of Preservation Strategy	Tick
Retention of original hardware and application software	
Migration of data	
Emulation	
Other (please specify)	

196. Have you adopted any quality control procedures in the life-cycle management?
Yes r No r

If yes, please specify (e.g. who checks and evaluates what, when, how, then what?)

197. For how long do you plan to keep these digital deliverables available?

198. Is longer-term sustainability dependent on self-generating funds?
Yes r No r

199. Will the resource generate sufficient income to sustain the resource?
Yes r No r

If no, has the project/program already secured the resources available for this?
Yes r No r

200. Has your project/program established an 'exit' strategy that could be adopted if your institution can no longer met the costs of sustaining the digital deliverables?
Yes r No r

If yes, can you describe that strategy?

201. Would loss of the digital deliverable matter?

202. Do you have any additional comments?

……..

……..

……..

Thank you very much for your time and help

II. Project Planning

Introduction

Planning is the first and arguably the most important step in any digitization project. Lest this sound like a platitude, it is worth noting that far too many projects are undertaken without adequate thought to the activities involved, the staff required, or the technical exigencies of the work. The need for good planning may be self-evident, but in practice it is often difficult to anticipate all the areas in which forethought is essential. Good planning for any project—even for managers who have successfully completed previous projects—requires a large number of decisions on questions such as the following:

- What work needs to be done;
- How it will be done (according to which standards, specifications, best practices);
- Who should do the work (and where);
- How long the work will take;
- How much it will cost, both to "resource" the infrastructure and to do the content conversion;
- Where, after having answered all of these questions, one might obtain funding.

This kind of planning is one of the most intellectually challenging of the project tasks, and may well be time-consuming. There may also be pressure to hurry this step, from a desire to show visible progress or in response to institutional pressure. But an investment in this kind of planning will be amply repaid over the life of the project: in the quality of the products, in smooth workflow, in staff morale, and not least in the total project cost. The goal of this section is to sketch out the parts of the planning process and indicate the important decisions—assessing the resources needed to complete the project, the staffing and equipment required, the choice and role of metadata, and the overall project management—and how to go about making them effectively. The checklist below gives a brief inventory of the resources required to undertake a digitization project. Not all projects will require all the resources listed, but this list will show the range of needs you should anticipate.

Technology develops and changes so quickly that decisions like those listed above may seem almost impossible to make with any confidence. Information on the array of standards, specifications, equipment, skills, and techniques not only presents a daunting learning curve, but also a welter of detail that can be very difficult to track. For the project planner, however, it is not these details that really inform good decision-making. It is much less important to know what sampling rate a particular piece of equipment offers than to understand how sampling works and how it can affect the quality of digital conversion. These underlying principles apply more broadly and change more slowly. Most importantly, though, they represent the level at which good planning takes place; with this knowledge, the planner has the tools to bring together an expert group of staff and consultants and create an effective framework within which they can work. This *Guide* contains detailed, up-to-date information on best practices in a number of technical areas, but the *Guide's* greatest and most enduring value for the project planner is its presentation of the more fundamental issues and how they interrelate.

The *Guide's* introductory section has already addressed the first question on the list above: What work needs to be done? By emphasizing the identification of audience and of your own institutional location and goals, the introduction contextualizes this decision and reminds us to ask "Who needs this work? Who will benefit?" The further ramifications of this question are explored in Section III on selecting materials, which discusses how to assess your collections and set priorities for digitization, and in Section XII on user evaluation, which provides guidance on how to assess the needs of your audience and how this information can shape your digitization strategy. This is also the stage at which you should get the facts and make your decisions concerning rights management, without which you cannot proceed with digitization: you need to establish the intellectual property status of the materials you wish to digitize, and you also need to decide on your own strategy for managing the intellectual property you are about to create. Both of these issues are explored in depth in Section IV. And although the project's final product may seem impossibly remote at this stage, you need to consider how the results will be distributed: not only what technologies you will use, but also how you will control access and ensure that you reach your intended audience. Section X covers these issues in detail.

The question of how the work will be done—the specifications, standards, and procedures you need to establish—has many facets which are addressed at various points in the *Guide*. Foremost among these is the question of standards: by using standards-based approaches wherever possible, you increase the longevity, portability, and interoperability of your data. You need to be aware of the standards that apply to the kinds of digitization you are undertaking, and these are described in detail in the sections on digitizing text, images, and audio-visual materials. Given the complexity and breadth of most standards, though, you also need to be aware of the best practices that apply to your community. For instance, both documentary historians and linguistic researchers use the XML-based Text Encoding Initiative Guidelines to encode textual data, but each group uses the standard in different ways that serve their particular needs. While you are considering the specifications for your data, you should also think carefully about how to capture and represent the metadata you will need to administer your digital materials and enable them to be used effectively. The *Guide* includes an appendix on metadata which describes the various types and their uses. The relevant sections of the *Guide* also provide pointers to specific information on best practices for particular digitization communities.

The question of "how" also involves decisions about equipment. For the project planner, these questions are most usefully addressed not at the level of specific brands and models, but by thinking about the functionality you require and the tradeoffs you are willing to make (for instance, whether keeping costs low is more important to the project's overall success than achieving the highest possible capture standard). The sections on images and audio-visual materials discuss how to approach these decisions; more specific information on particular kinds of equipment can be found in the appendix on equipment. Finally, you need to establish an effective workflow for your project. At the highest level, this includes project management strategies, which are discussed later in this section, and quality assurance methods (discussed in Section VIII). But in addition you need to consider how you will store, manage, and track your digital objects, which is addressed in detail in Section XIII on digital asset management.

Staffing issues—who should do the work—are closely related to the points just mentioned, since your decisions about methods and procedures may be difficult to separate from the staff resources you actually have available. Few projects have the luxury of hiring all new staff to an arbitrary standard of skill and experience. Further on in this section we discuss human resources: how to construct job descriptions and identify skilled staff, and how to set up a management and advisory framework that allows your staff the autonomy to do their jobs effectively. In Section IX, Working With Others, we consider a range of collaborative and cooperative relationships that may expand your staffing options, including project consultants, vendor outsourcing, collaboration with other institutions, and internal cooperation.

Once you have worked through the issues sketched above, you will be in a position to assess the practical scope of the project: how long the work will take, and how much it will cost. Of all the questions addressed here, these may be the most vulnerable to change over time, as techniques and equipment improve and grow cheaper, and as quality expectations rise. Some guidance on cost estimation is offered later in this section, and also in the sections on specific digitization areas (Sections V, VI, and VII). You should make sure in researching costs to take into account all of the startup and infrastructural costs the project will incur-costs for initial planning, choosing data specifications, building or choosing tracking and documentation systems, training staff, and so forth-as well as the incremental cost of digitizing the materials themselves. This is also an opportunity to consider the scope of your investment and whether this infrastructure can be reused or adapted for further digitization projects once this project is completed.

Finding the funds to undertake the project is the final step, at least logically; a successful funding request will almost always require a thorough consideration of the issues just described. Even if you are fortunate enough to have funding already committed, going through this process will ensure that you spend your resources prudently and receive value for your investment. Funding sources and strategies are discussed later in this section, and also in Section XI on sustainability.

The checklist box below gives a condensed list of the resources you may need to undertake a digitization project. Although not all projects will need all of the resources listed, it gives a sense of the range and options.

Checklist Box:

Resources that you will need for a digitization project:

Personnel:	advisors project management staff rights specialists researchers editors authors digitizers catalogers technical support/development legal advisors	☐
Software:	operating systems applications: --> image manipulation --> metadata authoring --> database --> indexing and search engine --> web server utilities server systems network clients specialist applications/developments	☐
Storage devices:	local hard drives network storage servers optical devices (e.g. CD writers) magnetic devices (e.g. tape drives) controlled storage environment	☐
Network infrastructure:	cables routers switches network cards ports	☐
Consumables:	stationery utilities printer cartridges lamps (for capture devices/special lighting) storage and backup media	☐
Project management:	preparing bids recruitment publicity and dissemination creation of deliverable product specifications design of workflow supervision of staff quality assurance	☐

Resources within your institution

If you are working within an institution that has other digitization projects under way, an examination of the resources already available within your institution is a good starting point. Staff will know if their department or unit has capture devices available or workers with experience of digitization or cataloging. This is an easy first step towards building a resource inventory, although knowing that you have one flatbed scanner, a digital camera and suitable equipment for digitizing audio, as well as people who know how to use that equipment, is

not on its own sufficient. A thorough identification of internal resources involves checking that:

- equipment and software are of a sufficient specification to meet your requirements;
- workers who can operate the equipment are available and appropriately trained;
- technical support and maintenance are in place;
- capture devices are (or can be) directly connected to your storage area; and,
- access to equipment and staff suits your workflow requirements.

Clearly assessing the adequacy of these resources is predicated on other decisions, such as your workflow requirements; indeed, many of the planning areas discussed in this section are closely interdependent. It should also be apparent why the *Guide's* introductory section stressed early on that you need to define what you want to do and the audience or audiences you intend to reach (see Section I). A clear statement of objectives (preferably in a formal document that can be shared with staff), combined with the resource inventory, will enable you to assess the suitability of your local resources.

You will make this document an even more effective planning tool by adding information about equipment specification (e.g. computer processor speed, RAM, hard disk capacity) and the results of tests for suitability. Before you can conclude that you have suitable resources you must test them to make certain that they will meet the requirements of the project. The Example Box below, "Resource Inventory and Test", shows what a resource inventory and test for scanners might look like.

Example Box:

Resource Inventory and Test:

PCs and Scanners	Functional Requirements	Suitability Test Result
1 Pentium 3, 600 Mhz, 128 MB Ram	Must handle processing and manipulation of image files up to 50 MB	Needs more RAM
1 Pentium 4, 1 Ghz, 384 MB Ram		***Okay***
1 Agfa Arcus		***Okay***
1 Agfa DuoScan 1200		Transparency tray inadequate

Overall Conclusion: Upgrade one PC and replace one scanner

Most large institutions in the cultural heritage sector will have resources that may be useful to the project but would not necessarily need to be borrowed for the entire life of the project. There may be physical equipment, such as devices for digital capture, analog capture equipment (e.g. record, tape, CD and video players that can be used when converting from analog to digital), network storage devices, or handling equipment and controlled storage for analog material.

Human resources may be even more useful—expertise in digitization, text encoding, networks or web delivery can often be found in-house. Even those institutions yet to carry out any significant digitization will have cognate areas of expertise. These skilled individuals can be difficult to find, so tell your colleagues that you are planning a digitization project and have them consider which skills might be of value to you. For example, the skills, techniques and processes required by digital photography are identical in many areas to analog photography, and the same applies to image processing. Similarly, the standards and methods for creating metadata have their roots in the creation of bibliographic records, library catalogs or finding aids and museum collection

management systems. In addition to this, it is important to consider the project team and project management process here. Projects should establish a set of procedures for project management from the very start of any project, identifying goals and time scales as well as tasks and outcomes tied to the availability of specific staff and equipment.

It is much easier to identify potential facilities and expertise within the framework of an institutional digitization policy or corporate technology plan—follow the more detailed questions for your own resources as described above. If such a policy has not already been adopted, it will probably be beyond the scope of an individual project to initiate one. Nevertheless, informal inquiries can still be made relatively easily. Remember that apparently unrelated departments or projects may be useful. For example, a great deal of high-end digital imaging takes place in dental, medical, biological and life science departments. The Internal Resource Identification Question Box illustrates some of the common areas of expertise to be found within an institution.

- Is there an institutional digitization policy to adhere to?
- Who else in the institution has digitization projects underway?
- What experience can you use (e.g., photographic, equipment analysis, etc.)?

Question Box:

Internal Resource Identification:

	Institution Type		
Resource	Academic	Library	Museum/Gallery
Imaging	Medical Imaging / Media Services / Photographic Services / Library	Special Collections / Photographic Dept	Imaging / Publications Dept
Metadata	Library	Cataloging Finding Aids	Collection Management Finding Aids
Text Encoding	Literature / Language / Computing Science Depts. / Information Management / Library	Cataloging / Information Management Finding Aids Electronic Texts	Finding Aids / Information Management

External Resources

Identifying resources outside your immediate department, unit or institution can be a more difficult process. Success depends upon what type of institution you are, your strengths and limitations, the accessibility of the resources you are seeking, and whether there is scope for collaboration. Guidance from and access to the experience of others are likely to be readily available. The Link Box points you to national organizations that provide information to support digitization projects. Outsourcing can be another way to fill gaps in the resources available locally, by contracting with a vendor, hiring a consultant, or establishing a cooperative relationship with another institution. These options are discussed in greater detail in Section IX, Working with Others.

Link Box:

Links to National Organizations Offering Guidance

CLIR: Council on Library and Information Resources: "The projects and activities of CLIR are aimed at ensuring that information resources needed by scholars, students, and the general public are available for future generations." http://www.clir.org/

DLIB Forum: "The D-Lib Forum supports the community of researchers and developers working to create and apply the technologies leading to the global digital library." http://www.dlib.org/

LOC: Library of Congress: "The Library's mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations." http://www.loc.gov/

NINCH: National Initiative for a Network Cultural Heritage: "A coalition of arts, humanities and social science organizations created to assure leadership from the cultural community in the evolution of the digital environment." http://www.ninch.org/

RLG: Research Libraries Group: "The Research Libraries Group, Inc., is a not-for-profit membership corporation of universities, archives, historical societies, museums, and other institutions devoted to improving access to information that supports research and learning." http://www.rlg.org/rlg.html

PADI: "The National Library of Australia's Preserving Access to Digital Information initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access." http://www.nla.gov.au/padi/

AHDS: Arts and Humanities Data Service: "Create and preserve digital collections in all areas of the arts and humanities." http://ahds.ac.uk/

HEDS: Higher Education Digitization Service: "The Service provides advice, consultancy and a complete production service for digitization and digital library development." http://heds.herts.ac.uk/

TASI: Technical Advisory Service for Images: "Advise and support the academic community on the digital creation, storage and delivery of image-related information." http://www.tasi.ac.uk/

Resource challenges

There are a number of challenges both in assessing and securing the resources required for the project. Projects that take place in large institutions frequently benefit from a significant amount of non-project-related investment. Such hidden benefits include local area networks, high bandwidth Internet connections, large capacity network-based storage devices, web servers, and technical expertise associated with maintaining and developing these facilities. This infrastructure provides the framework for the specific resources and skills a project needs, and without it many projects simply would never get off the ground. Although institutions are now trying to quantify this input, its actual value is difficult to establish, with the result that projects in well-resourced institutions are able to scale up more quickly but often under-represent the real costs that lie behind the their activities.

Equally, less well-resourced institutions and initiatives face an increasing challenge in matching the developments in presentation and delivery of digital resources that larger projects can provide. Frequently, the solution is for small and medium size institutions to develop collaborative projects. The Colorado Digitization Project (http://coloradodigital.coalliance.org/) provides a flagship example of how equipment, staff and expertise can be shared between large and small projects alike, enabling the digitization and delivery of resources that would not otherwise be possible.

Another challenge for digitization projects, large and small, lies in the area of human resources. Content creation is a burgeoning field and although many Internet businesses have failed, those companies such as Getty Images, Corbis, The Wall Street Journal and Reed Elsevier, which have adopted prudent content creation and marketing strategies, are showing steady growth. The finance, commerce, media and entertainment industries all recognize the value and benefits of digital assets, and this places a premium on skilled personnel. Furthermore, the development of staff with digitization skills related specifically to the humanities and cultural field has not kept pace with the growth in the number of digitization projects. Many projects report difficulties in recruiting and retaining staff. Few public sector projects can match the remuneration levels offered by the private sector, but there are strategies you can adopt that enhance your chances of meeting the human resources

challenge. These are outlined in the Human Resources Question Box.

Question Box:

Human Resources:

- Are there non-monetary factors that can be emphasized or enhanced? For example, will the project offer advantageous working conditions, training opportunities, or the possibility of gaining qualifications or accreditations?
- Are there aspects of the job that are more attractive than private sector equivalents (e.g. greater creativity, responsibility, freedom)?
- Can posts be combined or split to make most effective use of existing skills?
- Can you consider applicants from a non-humanities/cultural background, particularly for technical posts?
- Can any staff be re-deployed, temporarily transferred or re-trained from elsewhere in your institution?
- Can posts be shared or joint funded with other projects?
- Are you able to outsource any jobs?

Funding

Some project staff will be preoccupied with securing adequate financial resources to start, develop and sustain a project throughout its lifecycle. An accurate picture of the financial costs will help you to identify the financial pressure points and to estimate more accurately the overall costs of running the project. The sections below on skills, equipment, and project management will provide points to help you develop accurate project budgets. An accurate profile of project costs helps to minimize the financial unpredictability of the project and improves the probability that it will attract funding. Funding agencies remain attracted by the opportunities for funding initiatives in the heritage sector. The Link Box provides pointers to some major US funders.

Link Box:

Potential Funders of Digitization Projects:

- Andrew Mellon Foundation: The purpose of the Foundation is to "aid and promote such religious, charitable, scientific, literary, and educational purposes as may be in the furtherance of the public welfare or tend to promote the well-doing or well-being of mankind." http://www.mellon.org/awmf.html
- NEH: National Endowment for the Humanities, "an independent grant-making agency of the United States government dedicated to supporting research, education, and public programs in the humanities." http://www.neh.gov/
- The Getty: "The Getty Grant Program provides support to institutions and individuals throughout the world for projects that promote the understanding of art and its history and the conservation of cultural heritage." http://www.getty.edu/grants/
- IMLS: Institute of Museum and Library Services, "an independent federal agency that fosters leadership, innovation, and a lifetime of learning." http://www.imls.gov/grants/index.htm
- NHPRC: National Historical Publications and Records Commission, "supports a wide range of activities to preserve, publish, and encourage the use of documentary sources relating to the history of the United States." http://www.nara.gov/nhprc/

From the projects surveyed it is evident that most potential funders, particularly in the public sector, require applicants to provide a robust and auditable cost model. How this should be presented may vary from one funder to another, but it can be extremely useful to break down equipment and salary costs on a per unit or work package basis. Not only does it help the potential funders to make comparisons of unit costs between

projects within and across heritage sectors, but it also forces you to look at the process and scheduling of work in detail. The accuracy of these figures will be greatly improved by conducting a pilot study or by adopting a cost model from a previous project, even if it needs to be revised in light of the experience of the earlier project.

All the projects surveyed obtained their financial backing from a combination of institutional budgets, public grants, private donation or corporate sponsorship. None of the projects reported serious under-funding, although some found that the distribution of funds created an uneven cash flow, resulting in medium term planning problems. Similarly, none of the projects reported serious concerns about sustainability, even where the source of future funds was unclear. The general absence of plans for self-generating funds or of exit strategies supports this confident view that income would continue to materialize in the future. A number of projects have recognized that failing to adopt long-term financial planning is less than prudent. We recommend that time and support for securing further external funds are crucial as well as exploring the potential for self-generating income. Projects should develop an exit strategy that will secure the maintenance and accessibility of the digital material. These issues are discussed in more detail in Section XI on Sustainability.

Cost models

Determining the cost of digital content creation on a per unit basis is extremely problematic. Not only are there no comprehensive cost models available that cover all resource types but trying to apply such a model to the variety of institution types, financial arrangements, prevailing market conditions, nature and volume of material and the resolutions required would be problematic. Furthermore, the cost basis for creating, storing and delivering digital resources can be quite different and trying to establish a single cost per unit can disguise these differences or ignore them altogether. In spite of these problems it is possible to establish some bases for per unit cost.

At the simplest level a project can take the total funding required and divide it by the total number of units that they intend to digitize. For example total project funding of $300,000 divided by 40,000 units equals $7.5 per unit. However, such a figure can be extremely misleading. Firstly, there will be variation in per unit cost according to the type of material digitized. The creation of OCR text pages will differ from reflective color still images, which will be different again from 16mm moving images or 78 rpm records. Even within material of the same broad type there will be variation. Black-and-white negatives are likely to be more expensive to scan than black-and-white prints, since tone reproduction needs to be set image-by-image in the former case, while the same settings can be applied to a group of photographic prints. Even if a project is dealing with material of a uniform medium and size, variations can occur that impact on unit costs. A collection of bound, legal-size books may have volumes that cannot be opened beyond a certain degree for conservation reasons. This may require a different capture technique, for example capturing pages from above rather than inverted. Some volumes may have details that demand a higher capture resolution than the rest of the collection, while others may require curatorial intervention to prepare them for digitization. The extent to which projects need to take account of such details will vary but at the very least different material types should be distinguished as well as same-type materials that require different capture techniques.

The cost items that go to make up a per unit calculation also require consideration. Should pre-digitization conservation work, handling time, programmers and management staff be included in addition to capture equipment and staff? In practice, projects need to do both. This is best achieved by calculating the costs directly related to capture on a per unit basis, which facilitates comparison and cost effectiveness for different techniques. Non-capture-related items could then be added to provide a total project cost and a second per unit calculation could be carried out if required. The list box below provides an indication of how these different factors can be differentiated. It is common practice to calculate costs for audio-visual material on a per minute basis.

List Box:

Capture Cost Factors:

(per unit for a single media type with uniform capture techniques and settings). It is important to note that the digitization capture costs are actually the least costly of the whole process.

- Handling time (from the shelf to point of capture and return) as a percentage of total salary costs on a

daily basis
- Pre-digitization conservation work (this should only be applied for those items that have required it)
- Capture time (from set-up to naming and saving) provided as a percentage of the capture operators total salary costs on a daily basis
- Cataloging/Metadata (required for digitization and/or created at capture stage) as a percentage of total salary costs
- Hardware cost per item
- Quality Assurance time as a percentage of salary cost
- Software cost per item (both hardware and software costs should be on the basis of the depreciation of equipment or projected replacement cost, rather than the total cost of hardware and software)
- Hardware maintenance
- Technical support time (proportion of total salary or contract cost related to capture)
- Project Management time (proportion of total salary related to capture)
- Training (directly related to capture)

Non-Capture Cost Factors:

- Additional Project Management salary
- Web Programmer's salary
- Educational Officer's salary (or other additional project staff)
- Cataloging/Metadata (post capture creation) % of total salary costs
- Additional technical support salary
- Additional hardware and software costs
- Consumables (including any storage media such as CDs or DATs)
- Travel and per diem
- Training (non-capture related)
- Storage costs (based on total maintained cost for the gigabytes required)

Some sites with detailed information on costing are listed below.

Key Sites with resources on costings:

- Research Libraries Group: http://www.rlg.org
- Library of Congress: http://www.loc.gov
- Online Computer Library Center: http://www.oclc.org/home/

Human Resources

A project's long-term success depends on the accurate assessment of the required human resources, and producing a map of available and unavailable skills is a valuable starting point. Institutions vary in their areas of expertise and different types of project require different skills. Nevertheless, from the projects that we surveyed it has proved possible to develop a basic template of the people and skills required in realizing a digitization project. The requirements can be scaled according to the size of the project envisaged.

Job descriptions, performance indicators, training

Comprehensive job descriptions are indispensable, regardless of the project or institution. While job descriptions are not always required by the host institution, employment law often demands them. Funders are increasingly expressing an interest in viewing job descriptions as part of the application process as this provides them with a richer overview of the project. It is worthwhile developing an outline of job descriptions before the project reaches the recruitment stage. This is useful to determine the delegation of work, how jobs interrelate, which posts can be tailored to existing skills and which can be identified for external recruitment or outsourcing. A

useful process for developing accurate job descriptions is to set out a list of all the tasks required for a post and then rank them from highest to lowest priority or into essential, desirable and non-essential categories. Next, compile a corresponding list linking these tasks to the skills required, including any particular knowledge or qualification. Alongside this, compose a description of the experience or background required for these skills. Finally, review the original tasks and their priority to ensure that a realistic and coherent job description is produced. A resource which has been developed by the Association for Computers and the Humanities is a database of jobs in this field—it may be consulted by projects for guidance in drafting job descriptions, and can also be used to publicize new jobs to a focused audience of candidates. See http://www.ach.org/jobs/ for more information.

Example Box:

Sample Job Description

Job title: Digital Library Research Assistant

The Digital Library Research Assistant will play an integral role in the university's digital library projects, the goal of which is to bring a wide range of source materials to as large an audience as possible. The DLRA has responsibility for overseeing initial scanning and data capture, creating and reviewing metadata, and performing quality assurance checks. With other project members, collaborates on project publications and research.

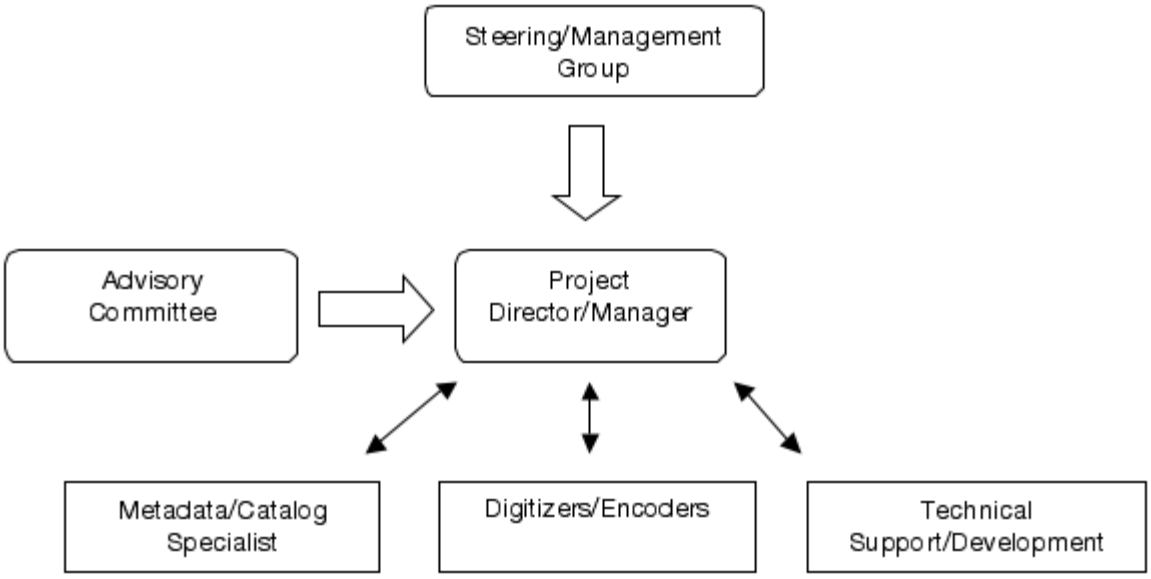
Job requirements: Bachelor's degree and one to three years' experience; basic computational skills, and expertise in at least one area of the humanities. Advanced degree and three to five years experience preferred. Familiarity with relevant encoding and metadata standards, including SGML/XML, METS and Dublin Core, is highly desirable. Must be a self-directed team worker with strong motivation and the ability to take initiative. Needs good communications skills (oral and written) and willingness to work collaboratively.

The use of performance indicators appears to be on the increase. They can have a positive impact, not least by providing a way of formally identifying training requirements. While most projects assess training needs on the job as an informal exercise, formal methods encourage appropriate training solutions to be planned and resourced in advance.

There is a close interplay between performance indicators, job descriptions and training assessments. The job description is very useful in developing meaningful performance indicators. Indeed, a useful starting point for performance review is to evaluate current tasks against those set out in the job description, highlighting whether the original job description was unrealistic, whether workloads need to be re-evaluated in the light of practical experience, or whether a skills shortfall needs to be addressed. The aim of addressing training requirements is to ensure that future tasks can be achieved and that the project will not encounter a skill shortage.

Managing the skills base

It is vital to ensure that a project be able draw on the right balance of skills. The challenge is to determine the skills of individuals and how they can most effectively contribute to the project. The key to successful delivery of projects is management. The diagram below incorporates elements from all of the projects surveyed, from the smallest to the largest, and illustrates the general structure that may be used to manage the project's skills base.



The steering group functions as an executive board and includes all constituents who are directly involved in the project, even if not employed by it, such as curators, archivists, subject specialists and education officers. In practice it is common for the steering group to be an existing committee within an institution.

The advisory committee is a broader-based group, providing general advice on the project's focus and direction. Members usually include the steering group with additional appointments from external organizations bringing particular areas of expertise, such as evaluation, to the initiative. There may be more than one advisory committee, or the advisory committee may be broken down into sub-committees each of which supplies more focused technical, academic or editorial decision-making support. This is the case with the Perseus Project at Tufts University, which has separate Technical and Academic Advisory Boards as well as a Steering Group to provide general project management. (Read Interview 28.2 for details on this arrangement)

It is essential to have a single project manager who is employed by the project, with responsibility for its daily management. In most cases the project manager provides the necessary project management experience, supplemented by internal or external advice. An institution needs to assign both accountability and authority to the project manager position, so that the process is not bogged down by myriad interactions with the advisory group or groups to deal with daily operations. In content creation projects it is unusual to employ external consultants to handle project management.

What skills are required?

There are four main areas, which will require staff with identifiable skills. These skill areas may be provided within a single project, dispersed across a collaborative project, or outsourced.

- Conservation: A crucial aspect of any digitization initiative will be a conservation assessment of the analog materials. Under some conditions this may show that before some material can be digitized it will require conservation intervention.
- Digitization/Encoding: This can involve digital imaging, keyboarding, OCR, character or full-text encoding, or a combination of these. In some projects it may also include conservation intervention in the analog material.
- Metadata/Cataloging: The creation of metadata records for the digital material. This work may also involve cataloging the analog material or searching for information to enhance the metadata record where it is absent from the analog version.
- Technical Development/Support: This falls into two distinct areas: the creation or implementation of specific IT solutions for creating, managing or delivering the digital material, and the provision of IT support for project hardware and software. This latter area includes desktop applications, network services, and capture devices.

In smaller projects staff may carry out tasks in more than one area: for example, the digitizer may also undertake technical development, or the project manager may take on metadata creation. In larger projects, such as SHOAH or the Genealogical Society of Utah, the duties of staff are so extensive that this is not feasible.

Project managers will have to decide whether to hire new staff with the required skills or to re-deploy existing staff from other areas of the institution. We found that many projects prefer the former, with two notable exceptions. First, there is a discernable trend for photographers to be employed for high-end digitization work. Projects have found that better-quality images are produced through training a photographer in digitization rather than trying to equip a digitizer with photographic skills. The second exception is the tendency to re-deploy or train existing cataloging staff in metadata creation. This is a logical progression for staff who will already have considerable experience in creating bibliographic records, collection management records, finding aids or catalogs, frequently in an electronic form such as MARC.

Another decision concerns background skills. With the exception of some technical posts, we noted a clear preference for staff with arts, humanities, library, museum or gallery backgrounds, or at least some experience or interest in the subject area of the collection. There may sometimes be advantages in not having such a specialization. For keyed-in text transcription, staff without subject knowledge are more likely to enter exactly what is on the page rather than interpret the contents and enter what they think is in text. On the other hand, subject knowledge can be exceptionally useful in gauging what areas of the content should be focused upon, deciphering difficult materials, or recognizing how areas of the content should be marked up.

When you are trying to find staff with appropriate skills, remember that some projects have benefited from using student labor and volunteers. The ability to draw on student labor represents a significant benefit for university-based projects. Projects such as those based at the University of Virginia Library have been able to build large and diverse digital collections because they are able to draw upon a pool of skilled, motivated and affordable labor. Projects that recruit student labor have invested considerably in training, adopted flexible working practices and tailored the work around the students' educational commitments. This approach has the added benefit of equipping students with the skill set required for future work, adding to the pool of available staff.

Volunteers often provide a similar pool of skills and projects such as the Genealogical Society of Utah have made effective use of this resource. They have found it both necessary and beneficial to invest in appropriate training for the volunteers. Such training should be factored into the project resource plans. In large-scale initiatives, volunteer management and training may become a significant part of the project itself.

The Link Box below provides links to sites that support skills development in digital representation.

Link Box:

An increasing number of organizations are offering training in digitization, which generally proves cheaper and far more useful than commercial training courses:

- Archive Builders: Document Imaging and Document Management. http://www.ArchiveBuilders.com
- Cornell University Library, Department of Preservation and Conservation: Moving Theory into Practice: Digital Imaging for Libraries and Archives. "This workshop is intended for librarians, archivists, curators, administrators, technologists, and others who are contemplating or implementing digital imaging programs." http://www.library.cornell.edu/preservation/workshop/
- HATII: Humanities Advanced Technology and Information Institute. Digitization Summer School: "The course will examine the advantages of developing digital collections of heritage materials, as well as investigate issues involved in creating, curating, and managing access to such collections." http://www.hatii.arts.gla.ac.uk/
- Humanities Computing Unit in Oxford: Summer Seminars covering a range of topics: http://www.hcu.ox.ac.uk/
- School for Scanning - North East Document Conservation Center: Creating, Managing and Preserving Digital Assets. http://www.nedcc.org
- Harvard Extension School, Museum Studies Certification Program. Muse E130: Information Technology for Museums. http://www.extension.harvard.edu/2001-02/courses/muse.shtml
- TASI: Technical Advisory Service on Images. Training courses "aimed at those involved in digitization projects, those who wish to capture images and those who wish to use digital images in teaching and research." http://www.tasi.ac.uk/training/training1.html

- UCLA/Getty course: Museum, Libraries and Archives: Summer Institute for Knowledge Sharing. http://skipper.gseis.ucla.edu/orgs/gettysi/html/summer.html
- University of Virginia Library: Digital Media Lab Services tutorials and short courses on digital image, video, and audio capture and editing. http://www.lib.virginia.edu/clemons/RMC/digilab-services.html

Equipment

Because our digitization capabilities are so strongly tied to—and limited by—the developing equipment technology, it is tempting to feel that the available technology should motivate our digitization strategies. However, on the contrary, it is vital to base equipment requirements on the characteristics of the collection/s to be digitized and on project needs, and not the other way around.

Although there are significant cost savings associated with outsourcing work to "offshore" production bureaus in Asia, the Far East, Mexico, etc, in cases where unique materials or special collections materials are to be digitized it is important that digitization should take place as close to the original as possible. Hence many projects will need to confront the complex questions of equipment specification and selection. A detailed discussion of matching material properties to hardware and capture settings can be found in Section VII on audio-visual materials. There is also further information on equipment choices in the appendix on equipment. At the moment we will focus on the basic differences in equipment and the technologies employed in order that the correct type of equipment resource can be procured for a project. Selecting the most appropriate equipment can be time consuming, but projects should not be deterred by the plethora of manufacturers and their competing claims. For example, the SCAN project (Scottish Archive Network) was initially unable to find a commercially available digital camera that exactly matched their requirements. Instead, they sourced a camera custom-made to their exact specification. This level of exactitude may be out of reach—and unnecessary—for most projects, but it is worth remembering that one need not be entirely constrained by what is commercially available.

Principles of digital data capture

Although there is a variety of capture devices for different applications, whether you are digitizing images, text, audio, video or 3D objects, the operating principles are the same. All digital capture devices take a sample of the analog source material to create a digital surrogate. This sample is made up of two elements: the sample rate and the sample depth. The sample rate describes how frequently readings are taken of the analog material. For example, in a digital image this would be the resolution, or the frequency per unit of area: the number of pixels per inch, expressed as pixels per inch (ppi) or dots per inch (dpi). An image captured at 600 ppi would have had 360,000 samples recorded per square inch. Similarly, for audio-visual materials the sample rate is the frequency per unit of time at which the source material is sampled. The sample depth is the amount of information recorded at each sampling point. For example, a sample depth of 24-bits would capture 8 bits for each of the three color channels (red, green and blue) at every sample point. For a more detailed explanation of sampling, see the appendix on digital data capture and Section VII on Audio-Visual Materials.

Selecting equipment

The medium, format, size, and fragility of the original material are among the primary factors affecting equipment choice. For text documents, flatbed scanners are suitable for single leaf, regular sized documents, provided the material does not go beyond the scanner's maximum imaging area (usually up to approximately US Letter size), or is put at risk by "sandwiching" it in the scanner. Large format flatbed scanners and sheet-feed scanners can handle single leaf, oversized documents. However, sheet-feed scanners put material at greater risk than flatbed scanners as the originals are pulled through a set of rollers. Drum scanners, whose imaging area is usually from 8" x 10" to 20" x 25", and digital cameras can also be used for oversize material, but they are an expensive option compared to flatbed scanners.

Bound pages that cannot be disbound, and pages in bindings that cannot open fully to 180 degrees require flatbed scanners with a right angle, prism, or overhead capture array. Digital cameras, with appropriate easels, book rests and weights are a versatile option for bound material. Camera beds or mounts, lighting, lenses, and filters all add to the cost and complication but make digital cameras more versatile tools for capturing manuscripts, bound volumes, original works of art, prints, out-size material and artifacts.

To achieve the highest quality scans of transparent media (e.g. 35mm slides and negatives, 6x4 and large format transparencies and microfilm) specialist equipment such as slide and film scanners, microfilm scanners or drum scanners should be used. Some flatbed scanners, with a dual light source, can handle transparent media though they often lack the dynamic range comparable to that supported by transparency scanners. However, you will not achieve as high a quality image as you would with a dedicated film or slide scanner. These have an inherently higher resolution, appropriate for the small size of the original, hold the transparencies more closely and securely, and frequently have negative color compensation to correct color casts for different types of film.

Audio and moving image materials present their own problems for digital capture. Not only is there a variety of source formats, including wax cylinders, 33, 45 and 78 rpm records, 8-track and cassette tapes, two-inch and VHS video in PAL and NTSC formats, but it is often very difficult to obtain access to analog devices for playback and linkage is difficult.

Definition Box:

Audio-Visual Facilities:

- Audio capture card required for sound material or video capture card required for moving images
- Source devices, such as 78 rpm record players and tape players.
- Mechanism for connecting these devices digitization equipment
- Intermediary device, such as a DAT (capable of handling ASEBU and SPDIF digital audio) machine

Link Box:

There are a number of audio and video digitization projects that are just getting started:

RAI: http://www.rai.it/portale

BRAVA: Broadcast Restoration of Archives through Video Analysis
http://www.ina.fr/recherche/projets/encours/brava/

COLLATE: Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material
http://www.collate.de/index.htm

PRESTO: Preservation Technology for European Broadcast Archives http://presto.joanneum.ac.at/index.asp

AMICITIA: Asset Management Integration of Cultural Heritage In The Interchange between Archives
http://www.amicitia-project.de/ami_home.html

The 3D representation of objects, from coins to buildings, is at the forefront of current digitization developments. At present the technology can be divided into two broad categories. The first, and simplest, is to create a moving image of an object. This is achieved by moving a digital camera around the object, or rotating the object in front of a fixed camera, while taking a series of still images. These images are then compiled to create a moving image of the object. The most common format for this is QuickTime VR. This is a reliable technology that requires a digital camera and mount or turntable. However, it does not provide a true 3D representation of the object because while only two planes are captured and displayed, it still represents 3D objects using two spatial planes. The viewer cannot manipulate the object, and the views provided are fixed and pre-determined.

Creating a true 3D representation of an object requires that the dimensions and features of the object be modeled. That is, the three dimensions of the object are represented in the computer as a set of coordinates. Attached to this "frame" are the textures of the object to provide the surface details. At present most 3D imaging technology remains in the sphere of industry. The technologies used to capture coordinates, render the model, and interact with the 3D representation (such as haptic feedback systems that allow one to "touch" the object, or 3D printing to create facsimiles) are often quite costly and require a relatively enormous amount of computing

processor power compared to the average desktop computer (in 2002). As such, 3D modeling devices remain application-specific, for example body imaging, prototyping or CAD/CAM applications. However, it was not long ago that digital imaging was the sole preserve of medical applications. During the next ten years we should see increasingly cost-effective and user-friendly devices that will bring 3D modeling into the mainstream.

Definition Box:

Virtual Reality:

Virtual reality can be described as an interactive, self directed, multi-sensory, computer generated experience which gives the user an illusion of participating in a three dimensional environment, even if a synthetic one. For cultural and heritage institutions, this may mean using virtual reality to create virtual representations of three dimensional objects in their collections or to create representations of environments, such as an Egyptian tomb, an ancient Persian palace, a historic Greek theatre or an ancient landscape. These three-dimensional objects could range from coins, vases, and sculptures to representations of whole rooms of collections.

Metadata

Metadata is an indispensable part of any responsible digitization program, and considerable attention has been paid to the definition of high-quality metadata standards for various purposes. (The appendix on metadata provides more detail on different types of metadata, and on specific metadata schemes and their uses.)The availability of accurate metadata is as important as the digital surrogates themselves for accessibility, usability and effective asset management. In many instances institutions will already have substantial metadata about the analog object (for instance, catalog records) much of which can be applied to the digital object. The project will be able to reduce its metadata creation costs by building on existing metadata. When selecting material for digitization you may wish to give priority to material for which partial metadata already exists.

It is crucial to remember to determine the status of the existing metadata, when you are assessing resource requirements. In an ideal world the existing catalog or finding aid would be complete and up to date. However, many libraries, archives and museums have a backlog of cataloging work, and part of a collection selected for digitization could fall into this category. Therefore, it may be necessary to devote time to locating missing information for your metadata records. You must then decide whether to seek information just for those fields required for the metadata, or to update the original catalog record in its entirety. Digitization provides an economical opportunity for institutions to expand their metadata, so consider the possibility of seeking extra funds or devoting more resources to this activity. Some of the new elements required for the metadata record of the digital object can be generated automatically: for instance, automatic metadata creation is a feature of much high-end digital camera software and of some OCR systems. Alternatively, a project may need to develop its own system, and can greatly improve the efficiency and accuracy of technical metadata. There is a general dearth of metadata tools, which poses a problem for the efficient creation and management of metadata for many projects. There is therefore likely to be a significant element of manual work, whether this lies in adding digital objects to existing electronic catalogs, creating records for web-based delivery such as Dublin Core, or implementing encoded metadata schemes such as EAD. Creating a metadata record will usually take as long as creating the digital surrogate and if detailed encoding schemes such as Encoded Archival Description or Text Encoding Initiative are used, this process can be considerably longer.

METADATA RESOURCES:

GENERAL METADATA RESOURCES

1. Canadian Heritage Information Network Standards Page: http://www.chin.gc.ca/English/Standards/metadata_intro.html
2. J. Paul Getty Trust, Introduction to Metadata: http://www.getty.edu/research/institute/standards/intrometadata/

3. Extensible Markup Language: http://www.w3.org/XML/
4. International Federation of Library Associations and institutions. Digital Libraries: Metadata Resources: http://www.ifla.org/II/metadata.htm
5. Text Encoding Initiative: http://www.tei-c.org
6. Metadata Encoding and Transmission Standard (METS): http://www.loc.gov/standards/mets/

METADATA MENTIONED ELSEWHERE IN THE GUIDE

1. Section III: Selecting Materials: Metadata & Interoperability. The Dublin Core metadata initiative http://dublincore.org/
2. Section IV: Rights Management: Technologies for Copyright Management and Protection.
    - The Open Digital Rights Language Initiative (ODRL): http://odrl.net/
    - Digital Object Identifier (DOI): http://www.doi.org
3. Section V: Digitization and Encoding of Text - Text markup schema. Text Encoding Initiative (TEI): http://www.tei-c.org
4. Section VI: Images
    - Descriptive:
        - Library of Congress Subject Headings (LCSH): http://lcweb.loc.gov/cds/lcsh.html#lcsh20
        - Categories for the Description of Works of Art (CDWA): http://www.getty.edu/research/institute/standards/cdwa/
        - Art and Architecture Thesaurus (AAT) http://www.getty.edu/research/tools/vocabulary/aat/about.html
        - VRA Core Categories: http://www.vraweb.org/vracore3.htm
        - Dublin Core Metadata Element Set: http://dublincore.org/documents/dces/
    - Structural:
        - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
        - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
    - Administrative:
        - A Web Hub for Developing Administrative Metadata for Electronic Resource Management http://www.library.cornell.edu/cts/elicensestudy/
        - Digital Library Federation, "Structural, technical, and administrative metadata standards. A discussion document:" http://www.diglib.org/standards/stamdframe.htm
5. Section VII: Audio and Video Capture and Management
    - Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
    - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
    - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
    - MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
    - Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
    - Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
    - Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
    - Library of Congress's Digital Audio-Visual Extensions to METS Standard http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
    - Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html
6. Section VIII: Quality Control and Assurance: Importance of Quality Control and Assurance of Metadata
7. Section X: Distribution:
   Metadata Harvesting
    - Clifford Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/mhp.html
    - Donald Waters, "The Metadata Harvesting Initiative of the Mellon Foundation," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/waters.html
    - The OAI MHP protocol: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
    - MHP tutorial: http://library.cern.ch/HEPLW/4/papers/3/
    - CIMI Working Group: http://www.cimi.org/wg/metadata/
    - CLIR Metadata harvesting project: http://www.clir.org/activities/details/metadata-docs.html
    - DLF and Metadata Harvesting: http://www.diglib.org/architectures/mdharvest.htm
    - University of Illinois at Urbana-Champaign Metadata Harvesting services: http://oai.grainger.uiuc.edu/

8. Section XIII: Digital Asset Management: "Metadata definition and management"
9. Section XIV: Preservation:
   Institutional Approaches
   - METS (Metadata Encoding and Transmission Standard) http://www.loc.gov/standards/mets/
   - UK Cedars project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
   - Australian Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metadata: http://www.nla.gov.au/padi/topics/32.html
   - NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
   - OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
   - Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf

Project Management

Many different approaches to managing projects are possible. While we found little evidence of the conscious adoption of a project management model, such as PRINCE 2 (http://www.kay-uk.com/prince/princepm.htm), most projects implemented many of the key features of successful project management. As understanding of digitization becomes more commonplace it may not be necessary to "hot house" prototype projects in the manner that many early projects experienced. However, it should also be recognized that integrating existing projects into host institutions often adds a layer of bureaucracy.

The Genealogical Society of Utah provides a good example of a comprehensive project management model. Each imaging project undertaken follows six stages:

1. Negotiation and project administration
2. Capture Convert Acquire
3. Image and metadata processing
4. Storage and preservation
5. Indexing and cataloging
6. Access and distribution

All projects will need to consider these six areas in setting up their own project management systems.

You do not necessarily need to adopt all the activities of a project management methodology; rather you need to scale the method to the needs of your project. The whole process should be determined by the project's objectives and rationale for creating the digital deliverable. Each process should be defined, together with the specific objectives to be achieved and activities to be carried out. The various roles and responsibilities should be detailed (defining job descriptions and breaking finances down aid in this — see above) and adapted to the size and complexity of the project. This should enable the efficient control of resources and facilitate regular progress monitoring. Regular reviews should be used to ensure that the project's objectives, which may change during the project lifecycle, are being met. Whatever project management method is adopted, it should provide a common framework and delineate milestones for all elements of the project.

In summary, your project management methodology should make possible:

- The use of pilot projects and feasibility studies to shape the overall scheme of activity
- Controlled and organized stages
- The establishment of a project plan with milestones
- Regular reviews of progress against plan and against objectives
- Control of any deviations from the plan
- The involvement of all constituents at the right time and place during the project
- Good communication channels between all constituents in the project and the sponsoring institution/s

Other key features are the need for one project manager to have ultimate responsibility and for the project advisory group to provide management quality control and assurance. In distributed projects, site managers are recommended in addition to an overall project manager. Most projects have relied on internal project management expertise, supplemented by external advice. Although many projects started as relatively autonomous there is a clear trend for project management structures and the project organization to be

integrated into the host institution's structure. This may be a natural progression for projects as they mature, but new projects may consider whether they should adopt it immediately.

Work flow and costings

While few of the projects interviewed carried out benchmarking tests most had conducted pilot studies. These were undertaken for a variety of reasons:

- Technical feasibility
- Technical forecasting
- Workflow analysis
- Training needs

When considering technical forecasting or prototyping, particularly in relation to costs, remember that there may be no corresponding benefit, and if there is a benefit it will vary for different types of content. Few projects in the humanities and cultural sector charge users for the digital deliverables. As such the cost/benefit may simply be realized by the ability of the project to amortize the depreciation on the equipment. A new high-resolution camera may pay dividends for fine textual or line art material, but not so for color images. Similarly, a device that enables the digitization of material that previously could not be captured, such as a 3D modeler, may not make financial sense if a project has to build in a profit or depreciation margin. However, if the device makes an important collection more widely available, the public access benefit may outweigh the financial costs.

Where any form of pilot study is undertaken it is important to build this into the project design and development cycle. For example, the University of Virginia Library's Special Collections department delineates its project work as intricately as possible before extrapolating its workflow and costings. This has given the project reliable data to forecast costs, but there are some areas where measurement has proved inaccurate, such as network transfer rates. The UVA Special Collections department also has a scheduling calendar tied to a tracking database to generate quality control and assurance checks and back-ups. In this respect it is typical of the projects surveyed which all use flowcharts, spreadsheets or Gantt charts to plan and monitor their workflow and costs.

If you are considering using a cost model (see above), it is important to include all the relevant costs, not just the obvious items such as equipment and staff time. You will also need to decide on what basis to evaluate — for example, costs per unit to be digitized or costs per hour. The table below provides a checklist of the factors that should be built into a cost model.

Finally, one further area to be aware of as you develop your cost estimates is digital asset management. In digitizing an image collection, for instance, you may well be generating a number of different kinds of digital objects-archival masters, delivery masters, thumbnails and other deliverables-which in turn will require storage, tracking, documentation, and upkeep. This process may require a significant commitment of resources and will need to be planned carefully. Section XIII covers digital asset management in detail.

Cost Model Factors	
Equipment	Purchase Maintenance Repair
Software	Purchase Upgrades
Staff	Salary (including benefits and insurance,) Training Recruitment Travel & subsistence
	Heat Light

Utilities	Water Phone Postage
Building	Rates Maintenance Upgrading/expansion

Analysis Box:

Costs of Digitization Programs

There is little information available about costs and this is an area where future work is necessary. In a rare exception, Steven Puglia analyzed the costs of digitization programs, in particular from the Library of Congress Ameritech competition and the National Archives and Records Administration's Electronic Access Report. The costs discussed are mostly *projected* and *estimated* costs, a problem discussed in the conclusion, which suggests that further studies are necessary. After an initial discussion on general costs of projects—it appears that on average, a third of the costs incurred by projects is the digital conversion, slightly less than a third is metadata creation, and slightly more than a third is made up of administrative and quality assurance tasks—the emphasis turns towards long term maintenance costs. The author suggests that these are not often taken account of with the project costs.

Three types of maintenance of digital objects are considered, each with mounting costs in relation to the initial costs per image:

- The first projection is an example taken from the maintenance of NARA objects. This involves only keeping one copy of master files and offering thumbnails and access online. The cost of this basic maintenance is in the region of 14-38% of the initial imaging cost.
- The second projection comes from Cornell report on computer output microfilm. Applying this to NARA, the cost would be 55-60% of the initial cost per image.
- The third projection looks at outsourcing to a commercial firm. This would cost 275-796% of initial cost. But it must be remembered that the other two costs do not include major IT infrastructure costs and thus are false figures, whereas the private firm costing will include this.

In conclusion, it is suggested that digital imaging may not be the best approach for long-term retention of information. Institutions can only justify retention if the images are used. Analog retention is the best way of holding materials in the long-term. In addition, it would be instructive to use figures from final project costs and also examine costs per person and production per person.

For the full report see: Steven Puglia, 'The Costs of Digital Imaging Projects', *RLG DigiNews*, October 15 1999, Vol. 3 No. 5. http://www.rlg.org/preserv/diginews/diginews3-5.html

Conclusion

At the start of any project, project planning feels like a way to exert control, eliminate risk, and guarantee a successful outcome. Certainly without good planning, the likelihood of failure and inefficiency is much greater. But you can be a better project planner by recognizing that the goal is not to eliminate risk but to prepare for it—not to control every variable but to create a project framework within which your team's response to the unforeseen will be resourceful and effective. In the technology domain, change and unpredictability are facts of life, and often represent opportunities rather than disasters for a well-planned project. Your planning goal should be to create a flexible, adaptable system whose staff and procedures can accommodate change. Your aim as a project leader should be to distinguish between what is essential—the central project objectives, the strategic components that will ensure long-term viability—and what is merely instrumental detail.

valid xhtml 1.1

Ten rules for humanities scholars new to project management

derived from 15 years' experience as a digital humanities
PM/design editor/lead architect/PI and department director

1. Choose wisely. Do you really want to do this, and for how long? What contribution/important intervention is this project making?

2. Don't assume you need a grant. How much can you get done without one? (You'll be in better shape to ask for targeted support later.)

3. Seek partners, not services. Seek collaborators, not staff.

4. Draft a brief vision-document, make a mock-up, or otherwise set victory conditions. Iterate and refine with your partners, then freeze it.

5. Set clear internal milestones and issue tracking procedures ("This Project Powered by OCD") and state your big-picture deliverables/goals publicly (shame as PM's motivator).

6. What does sustainability mean to you/to this work? Secure stewardship agreements, embrace ephemerality as appropriate.

7. Foster community (users, developers) through small-scale publicity in just the right venues. Foster serendipity by communicating outside your comfort zone.

8. Enable decision-making on the part of every team member. Make clear and final decisions when needed. (Keep up momentum, keep up morale.)

9. Be the buffer for administrative and financial distractions.

10. Give all the credit away. But make it clear to your team that you'll take any blame.

Bethany Nowviskie
2012

Virtual Team Environment for Collaborative Research Projects

Lavinia Precup[1], David O'Sullivan[1*], Kathryn Cormican[1]and Lawrence Dooley[2]

1. *CIMRU, National University of Ireland, Nun's Island, Galway, Ireland*
2. *CEM, University of Dundee, Dundee, DD14HN, Scotland*
3.

** Corresponding Author: Tel: +35339750400; email: dos@nuigalway.ie*

Submitted to the OBMS International Journal of Applied Management Studies – 08-10-03

Abstract Large amounts of money are spent all over the world on research and development, in order to ensure future sustainability. However much of this research fails to meet its full potential. This escalates into bigger problems as research aims at global level projects that are needed to mobilise the knowledge, technological skills, money and experience required for today's complex products, services and processes. There is a wide scale of problems emulating from the management of such large-scale projects. These projects are often costly and inefficient especially when coupled with the problems associated with working in a virtual team environment, involving members from different countries and geographical regions. Problems surface when the knowledge of the many parties involved in the innovation project has to be shared and put to work to generate new knowledge and use it to create new products, processes and services. One of the main limiting factors that reduce the potential output of this type of project is the poor communication and knowledge sharing between members of the virtual team. In this context collaboration and virtual teams management are key issues that need to be addressed by today's innovation projects. This paper presents a literature review developed on the concept of collaboration between individuals working in a team and the advantages and requirements of virtual teams. Presenting a possible solution for the development of a virtual team environment that fosters innovation and ensures a good collaboration between individuals working in an innovation project. This solution was build using the Microsoft Sharepoint Services Beta 2 collaborative platform.

Keywords: Virtual Teams, Collaboration, Microsoft Sharepoint II, Distributed Innovation.

1. Introduction

Globalisation coupled with the instability that characterises the markets today and the technological explosion in the last few years that accelerated the rate of knowledge and information flow has put many organisations in difficulty to keep up to the extraordinary level and diversity of knowledge, experience and skills required to develop and successfully bring to market new services, products or processes. More and more companies are faced with the necessity to get the knowledge and expertise they require in different projects from different domains and areas. In many cases the knowledge necessary in development of these new products, services or processes do not fully reside inside the organisational boundaries. Consequently the 50-70's competition, become the 70-90's cooperation and evolved in the 21'st century collaboration. This transition from competition to collaboration in the development of new products raises one main problem: collaboration (Amidon, 2003). This problem comes from the fact that in order for the collaboration to be fructuous the people involved in the process, product or service development have to be open to share and acquire knowledge, skills, experience and expertise with other individuals. They have to be able to communicate at a high level with all the team members and use their knowledge and the knowledge acquired from others to come up with new ideas that can be put in practice or can become the germinating seed for other ideas.

Research has shown that 50-70 % of innovation project fail (Strebel, 1999). The causes for this high rate of failure are numerous and varied and poor collaboration between the team members has been identified as one of the key factors (Dooley et al., 2000). Poor collaboration, or in other words poor communication and knowledge sharing, between members of the project team highly reduces the potential output of the projects resulting in redundant information and knowledge and increased costs and lost of time. These problems

are accentuated when the team is a virtual team that consist of individuals that are located in different geographical areas, which come from different backgrounds and domains, different languages and cultures. When these issues are coupled with unstructured collaboration methods such as e-mails, conference calls and meetings, even if these methods are quick and familiar, new problems arise such as version control, work redundancy and maintenance of up-to-date information for all the members involved in the project. Various authors such as Nonaka and Takeuchi (Nonaka and Takeuchi, 1995), Davenport (Davenport and Prusak, 2000) and others identified and researched these knowledge management and collaboration problems. Different collaborative platforms have been developed in the last few years that strive to resolve the problem of collaboration between project team members whether they are located in the same or in different geographical areas from the same organisation or different organisations. Lotus Domino, eRoom, WelcomeHome, Microsoft SharePoint are just a few of the collaborative platforms that have developed to help organisations with the collaboration challenges. They answer in one way or another to the problems rose by collaboration and knowledge sharing. The success of these tools depends mainly on the level of involvement of the team members.

This paper presents, in the first part, a literature review that describes the concept and the types of collaboration and the dynamics of virtual team concept, which comprises of the virtual team definition, the advantages and disadvantages that come with the organisation of individuals in this type of team. The second part presents Microsoft Sharepoint Services as a platform for developing a virtual team environment and the virtual team environment developed using this platform for an innovation project.

2. Collaboration

To respond to competitive challenges, the complex customer requirements and the high instability on the market, firms have become more open to collaboration with other parties

than in the past. The creation of today's complex new products, processes or services requires a mixture of knowledge from diverse disciplines and professions and it is absolutely necessary to get the expertise from a variety of areas like information technology, technical design, engineering, manufacturing, and marketing etc. The experience, expertise and skills that are needed for an innovation project are usually distributed both within and outside a company. Therefore, people from different companies often need to work together to bring the entire knowledge and experience that are needed for the success of a new product, process or service. An effective knowledge management approach in these conditions can be the difference between success and failure of this type of inter-organisational collaborations. Knowledge sharing and transfer is not possible without willingness of employees to share and without a proper structure of the information to be shared. Tiwana (Tiwana, 2000) affirms that "knowledge based activities related to innovation and responsiveness are intensively collaborative", in other words the process of creating, sharing, and applying new knowledge and experience involves collaboration. According to Cormican and O'Sullivan (Cormican and O'Sullivan, 2001), a great deal of what people learn and implicitly what the organisation ends knowing results from interaction among and between the members of the teams that exist inside and between organisations. This underlines the importance of a good collaboration and communication between the members of projects teams that involve organisations that struggle to manage successfully their knowledge and innovation. The benefits of synchronized and collaborative working project team members are being recognised with team members being actively encouraged to work together more closely and to exchange project information in a more structured way (Anuba et al., 2002).

Based on the type of time and space the collaboration between the team members takes place Anuba et al (Anuba et al., 2001) identified four types of collaboration. Figure 1 depicts the space and time matrix for this collaboration types.

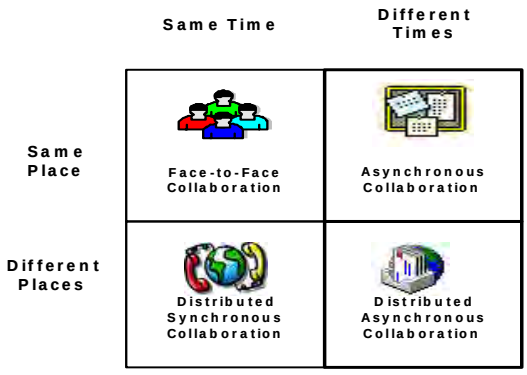


***Figure 1**: Collaboration types – Adapted from (Anuba et al., 2001)*

The four types of collaboration classified by Anuba (Anuba et al., 2001), based on the space and time the collaboration takes place, are Face-to-Face Collaboration, Asynchronous Collaboration, Distributed Synchronous Collaboration and Distributed Asynchronous Collaboration. The four types of collaboration will be presented in more detail next:

2.1. Face-to Face Collaboration

The face-to-face collaboration involves meetings in a common venue such as a meeting room, of the team members that are engaged in face-to-face discussions. An example could be a "kick-off" meeting at the beginning of a project between all the partners, and the discussions between the team members and the end customers for requirement gathering etc. This type of collaboration is the most common and requires good communication skills.

2.2. Asynchronous Collaboration

Asynchronous collaboration it is another type of collaboration between team members that has been identified by Anuba et al. (2001). The asynchronous collaboration process can be conducted using different mediums like notice/bulletin boards within an organisation or in a virtual environment that involves the same space/ environment but different times. In other words people can view the information, knowledge displayed at different times.

2.3. Distributed Synchronous Collaboration

The third type of collaboration identified by Anuba et al. (2001) as distributed synchronous collaboration involves communication between team members using the current technologies and techniques that imply the real-time exchange of information and knowledge between parties that are located in different geographical areas. There are various techniques available today such as telephony, video-telephony, computer-mediated conferencing, video conferencing, and electronic group discussions/editing facilities or on-line chats. This type of collaboration has become more and more used lately due to the extraordinary technological advancements and globalisation.

2.4. Distributed Asynchronous Collaboration

Finally, distributed asynchronous collaboration, the fourth type of collaboration involves collaboration between people that are located in different areas or places, which access the knowledge and information at different times. This mode of communication involves communication via the post, fax machines, telephone messages, voice mail, pagers, email, etc. The knowledge transmitted during this type of collaboration is an explicit knowledge that can be expressed in words, numbers and sentences.

A team, by its very nature, involves collaboration. In an organisation that facilitates collaboration the boundaries of the teams are usually flexible and unconstrained by the organization chart, department or even the company. In other words the team becomes an entity that is able to accommodate change without feeling any significant impact on its functionality, or that is able to quickly absorb the eventual impact. In a highly collaborative team the tasks are distributed to the team members based on their expertise, skills, and experience (Duffy, 1996). The refined blending of knowledge from different areas and domains required for the development of today's complex new products, processes or services required has changed the notion of the team involved in a project. In many cases, the participants in a project are geographically distributed, and need effective information and communication technologies in order to be always up to date with the information and knowledge in the project. In the last decade, a new term emerged, virtual team.

3. Virtual Team

Virtual teams bring up a different picture from the one of people in the same organization working together in the same place. Unlike traditional teams were all the members are located in the same organisation, department or building, and have the same working hours, share the same language and culture, the members of the virtual team work across space, time and organisational boundaries. They form webs of communication and can be from different geographic locations, different languages, cultures, time zones and even different organisations (Johnson et al., 2001). Different cultures, background domains, and research areas lead to issues that even an effective knowledge-sharing strategy might not resolve entirely. The virtual team members come from different disciplines and often "lack understanding of the critical process factors for areas other than their own" (Davenport et al., 2000). Contrasting with the traditional workspace, which is an office, department etc., the workplace of the virtual team is a "virtual workplace where productivity, flexibility and collaboration reach new levels" (Chase, 1999). With all these differences from the

traditional teams, new problem rise and old ones become critical. If the environment and collaboration were a problem in the traditional teams in the virtual team the need for a good collaboration and a proper environment for this collaboration are critical. To these problems are new ones are added such as contradictory physical space, geography, time, language, culture and knowledge sharing. For the virtual teams to be successful and have high performance results and a different kind of management of the team approach, an appropriate technological infrastructure and a well-developed environment are required.

In order to successfully and efficiently get all the experience needed in developing new products and services more and more organisations are forced to move from traditional face-to-face teams to virtual teams or adopt a combination between the two types of teams, lately. This shift of team management has its causes, mainly because of the advantages that come with the new team structure and the pressure place on the companies by the markets with the variety and increasing change rate of information and knowledge and the high competitiveness, by the customers that are becoming more and more demanding, and by the employees. Some of the basis for adopting a combination between a virtual team and a traditional team or for moving completing to a virtual team approach researched by Johnson et al (Johnson et al., 2001) are:

- Flexible hours for the employees – a high flexibility is given to the employees. The working hours are not strict and they are the one that are planning their working hours, as long as the team members do a good job and finish their work on time. Through this pressure and stress that is put on the employees working at a fix location following a strict timetable is reduced.
- Creates and disperses improved business processes across organisations – using the virtual team the best practices can be much easier presented and are quicker absorbed as the resistance to change is more diminished.

- Supports cross-divisional interaction – the virtual team structure of the team gives the companies the possibility to hire people that are from different domains and areas encouraging the team member in changing knowledge and information and creating the sense of a community crossing over the department, company or team boundaries
- Changes the direction of the company from a production-oriented company to a service/information-oriented company, as the focus is not as much on the product as on the team members' knowledge and ways to capture it to result in new, original ideas that will in the end bring profit. The new asset becomes the company's knowledge.
- Reduces costs for commutation between places by allowing the team members to interact regardless of their location.
- Allow organisations to hire and retain people, regardless of location
- Increased communication between team members, which leads to network building.
- Collaboration across organisational boundaries that facilitates the flow of information and knowledge between team members and constantly provides new ideas and new knowledge that can result in new products.
- Focus on measuring contributions and outcomes rather than on number of hours worked, as the knowledge and the original offspring generated by this knowledge are what is considered the asset in the new team structure.
- Faster response times to tasks as the result of each team member's work can be seen and evaluated by all the team members and because of that the sense of responsibility is more developed.

As the virtual team allows members to work together without requiring their physical presence in the same location, outside consultancy from anywhere in the world can be used, but without expenses for travel, lodging, and down time, partnerships between companies around the world can be realised to work on different projects without the relocation of the

team members. This way knowledge from many disciplines can be brought together along with the experience of the team members. While there are great advantages that come from the adoption of the virtual teams, new challenges rise with them. The main one involves satisfying the need for sharing, collaborating and exchanging information. There are some main requirements that have to be fulfilled so that the successful implementation of the virtual team can be realised. Some of these requirements researched by Lipnack and Stamps (1997) are:

- Collaborative efforts, e.g. co-authoring, document version controlling, applications development, group editing and reviewing, annotating, sharing information and ideas.
- Communication for e.g. sending notes, announces, talking, faxing, having meetings, providing commentary, and having interactive discussions.
- Co-ordination applications that keep track of communications and activities of workers in a group, preparation of group documents, and responses to questions received regardless of geography.
- Information sharing that helps users navigate the information rather than monitoring and controlling the communications patterns of the group.
- Managing and tracking, e.g. project management, corporate calendar, scheduler, version tracking.
- Analysis and decision making tools to support collaborative examination of information, consideration of possible decisions, and development of best decision.
- Developing collaboration between people who work at a distance from each other and who might not be from the same culture is a new organisation challenge (Pellecchia, 1998).

As discussed previously there are many advantages that come with the adoption of the virtual teams. The challenge is to provide an environment for the team were the team

members can communicate and efficiently manage the knowledge existing in the team and share it between the members of the team so that it will result in new ideas leading to new products, processes or services.

Due to the increased demand for a virtual team environment that will facilitate the collaboration between the members of the virtual team and will ensure the efficient exchange and management of knowledge in a project involving parties from different geographic areas, different domains and cultures a number of collaborative platforms and tools such as Microsoft Sharepoint, Lotus Domino, WelcomeHome emerged in the last few years. These tools have more or less common features such as discussion forums, instant messaging and chats, document storage, sharing and versioning, notice boards for announcements and notices and user administration features. This paper focuses on the Microsoft SharePoint Services as a collaborative platform for building a structured environment for a virtual team working in an innovation research and development project.

4. Building A Virtual Team Environment with Microsoft Sharepoint Services Beta 2

Microsoft Sharepoint Services is a virtual team website solution developed by Microsoft Corporation. It allows the development of a virtual team environment where the team members can store, share, and collaborate on-line on documents, contacts, tasks and knowledge necessary in the virtual team's efficient functioning. Different collaborative tools such as on-line discussions, announcements, notices and document sharing are available. These collaboration features can be extended as integration with Microsoft Outlook, MSN messenger and Net Meeting are available.

The platform allows the user to create a fully-featured website with build in functionalities such as announcements, meetings, discussion management, email notification and address book, survey, tasks and hyperlinks and also permits the development and creation of user-

defined functionalities. Other very important features, which address the interest of this research, include automatic hyperlink management, team management, and high team interaction. The website built using Microsoft SharePoint can be easily customized to meet the specific project demands using any web page editor such as Macromedia Dreamweaver MX, Microsoft Visual Studio Net or Microsoft FrontPage, which is fully integrated with Microsoft SharePoint, or other editors (Microsoft, 2002).

A virtual team environment was generated for a development and research innovation project using the Microsoft Sharepoint Services Beta 2 platform. The team members were distributed in different locations and they needed a working environment that would allow communicating and collaborating on different problems, ideas and documents regarding the project over the Internet.

By providing a user friendly interface and easy to use functionalities Sharepoint allowed the creation of the on line virtual team environment in a few easy to follow steps: fill in the name of the website, choose and website address, choose the website's type (virtual team environment, blank site, meeting workspace, document workspace, etc.), and finally chose a template. Figure 2 displays a screen shot of the Sharepoint web page for the creation of a new module/ page in the team website.

4.1.1 Sharepoint Build-In Features

A number of built-in features are implicitly generated when the SharePoint team site is created, such as Libraries, Lists, Discussion Boards, Surveys and Web Pages and will be presented in more detail next.

4.1.2 Libraries

A library is a collection of files that are shared between team members. There are tree types of libraries that can be created in SharePoint. These are document, picture, and form libraries. A built-in library named Shared Documents for storing, sharing and version tracking of the project's documents is implicitly created when the team environment is first generated (Microsoft, 2002).

4.1.3 Lists

A list is a collection of three preformatted forms module that allow the users to create, view, edit and delete different elements (entities) for storing and structuring project data information and knowledge.

A number of built-in lists are automatically generated when the virtual team environment which can be easily populated with project information and, if needed easily customisable the lists if desired. These lists are announcements (for sharing small amounts of information), links (for sharing of resources outside the team environment), events (for generating calendar-based view of meetings, deadlines, and other important events), contacts (for creating, sharing and managing information about the people that in a way or other interact with the team), issues (for creating lists of problems that appear), and tasks (for tracking the team members work progress). Additional lists can be created based on the built-in lists or based on the custom-designed lists. Once a list is created it can be saved as a custom-designed list and later be used as a base for generating new lists. Lists can also be created from spreadsheets from a compatible spreadsheet program such as Microsoft Excel 2002, and Microsoft Internet Explorer 5 or later browser. The hyperlink to a list in the website can be set to appear in the home page in the Quick Launch bar (Microsoft, 2002).

4.1.4 Discussion boards

Discussion boards offer the team members a forum where they can discuss or chat about topics of interest for the team. Comments and reply to others' comments can be posted eliminating the need for the e-mails and offering a knowledge centre for the other team members. A default discussion board named General Discussion is implicitly generated when the team environment is created (Microsoft, 2002).

4.1.5 Surveys

Surveys provide a method of finding team members opinion about different things of interest for the successful implementation of the project. To generate a survey, the survey's questions have to be defined together with the specifications on how the team members should enter their answers (Microsoft, 2002).

4.1.6 Web Pages

The Web Pages are used to add a new basic page, a web part page or a site under the team's website. A basic page can contain text, images and tables. A web part page provides a method for creating pages where web parts containing listings information from the website. Additional sites and wok spaces can be created under the team's web site such as document or discussions sites (Microsoft, 2002).

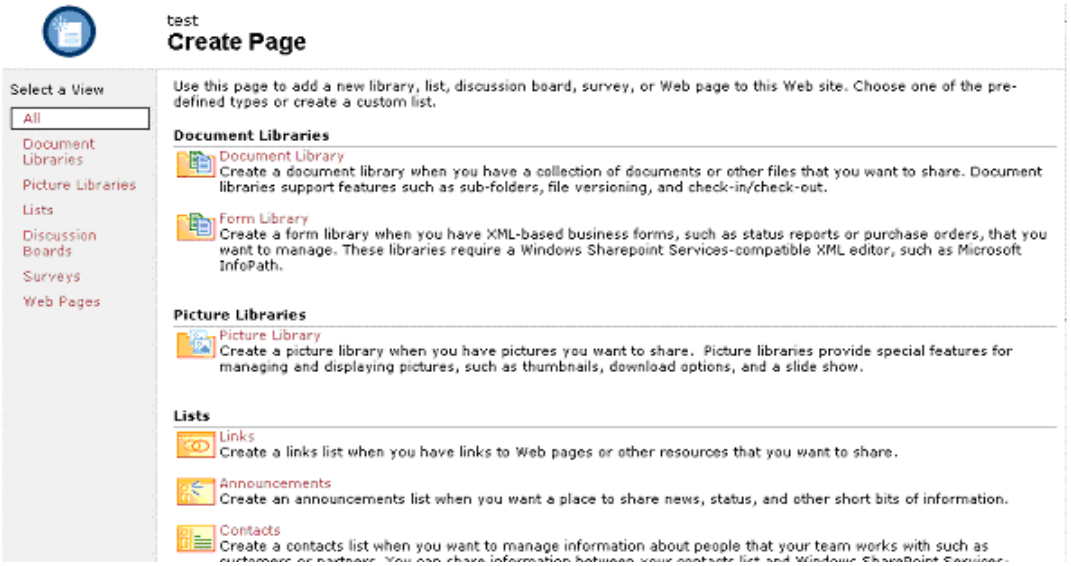


***Figure 2**: Create Page – Screen Shot*

4.2. Implicit Home Page Structure

A default home page for the virtual team is also generated when the new on line environment is created. The page is structured in four main areas:

- At the top of the home page a main toolbar - depicted in Figure 3 - with hyperlinks to the home page of the new virtual environment, the documents and lists/ modules in the environment, to the web page where new components can be added to the web site, to the site settings web page and to the parent website (if there is one) are displayed. Extra hyperlinks can be added edited or deleted from the toolbar using Microsoft FrontPage. This toolbar is implicitly displayed on all the pages in the website and can be removed using Microsoft FrontPage.

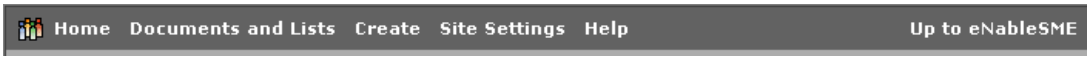


***Figure 3**: Main Toolbar – Screen Shot*

- Under the main toolbar an area with the website's name and brief description are displayed. The information displayed here is the one filled in when the new web site was created and it can be changed from the Site Settings web page
- A Quick Launch toolbar on the left hand side of the home page containing hyperlinks to the main modules in the website. The Quick Launch toolbar can be modified directly from the browser or from the web page editor. It implicitly displays the implicit modules
- A content area on the right hand side of the home page. This content area can be customised directly from the browser by dragging and dropping previously created web parts or from a web page editor such as Microsoft FrontPage.

4.3. Site Settings

The Site Settings web page depicted in Figure 4 is generated when the website is created and provides the authenticated team members with administrator rights to manage customise and change profile information and modify the structure of the site.

The administrators of the site can manage the users and the sites and workplaces that have been created on the website. The users can modify the title and the description of the site that appear on the first page under the main toolbar, to customise the home page and to modify the site content by adding new modules or modifying the created once. The user's profile can be modified from this page as well as depicted in Figure 4.

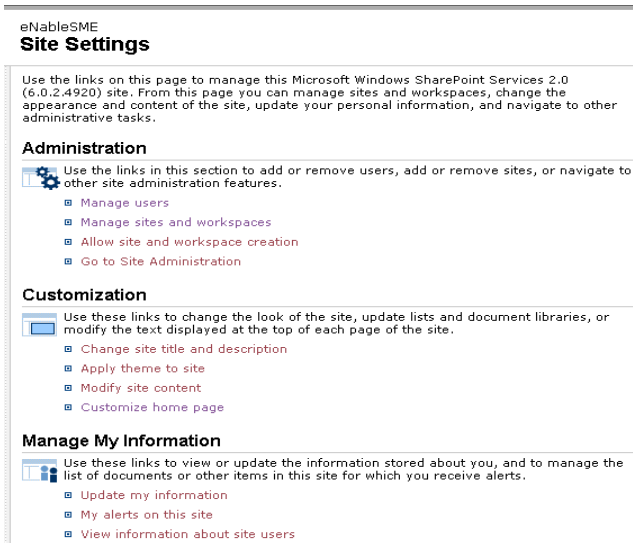


Figure 4: *Site Settings web page – Screen Shot*

4.3.1 Virtual team environment for project innovation management

Using the features provided by Sharepoint Services and the Microsoft FrontPage web page editor a virtual environment has been structured so that it facilitates the exchange of knowledge and the management of the innovation process in an innovation project. Figure 5 depicts the main page of the virtual team environment.

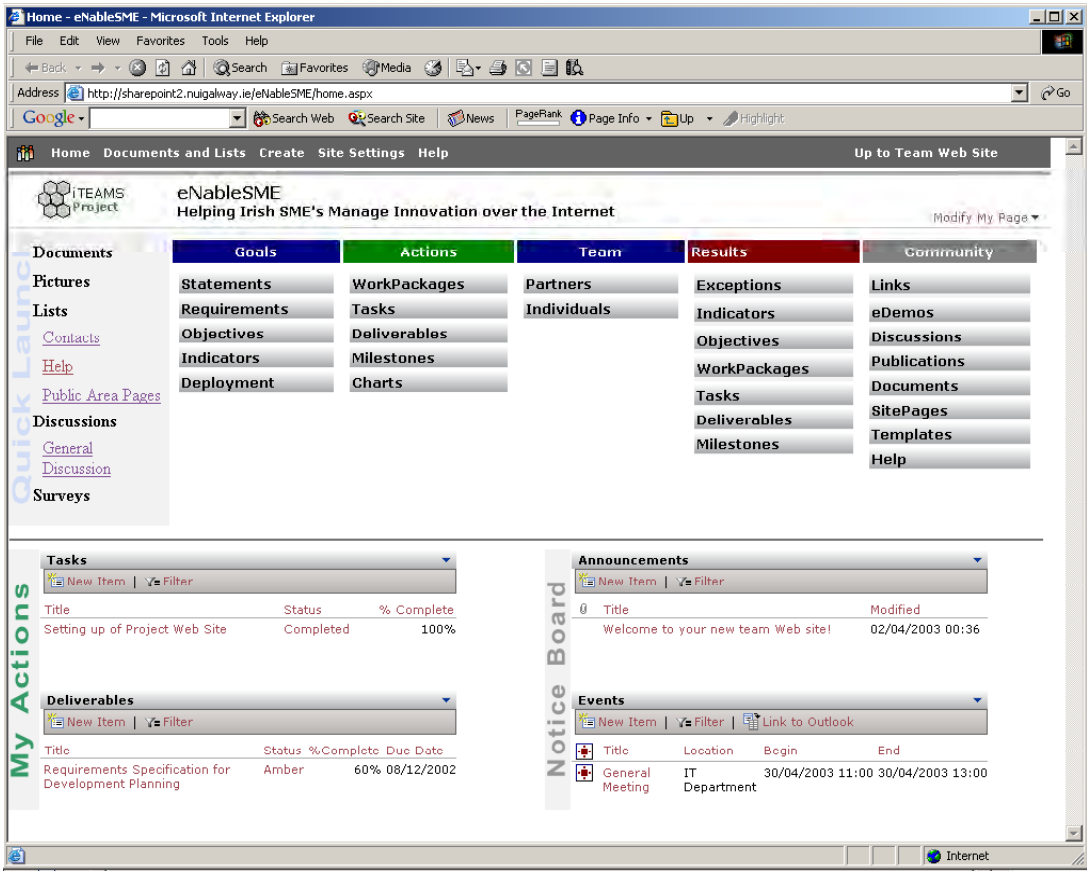


Figure 5: Virtual Team Environment Home Page – Screen Shot

The home page is the main page of the virtual environment and allows the users to access all the modules and sub-modules in the environments. The implicit main page generated when the virtual environment was generated has been customised so that it will provide the required environment for the management of project innovation process. The logo of the application, the name of the project and the project mission are displayed at the top of the page. The content of the Quick Launch toolbar was modified so it displays other hyperlinks as shown in Figure 5. The content of the home page has been modified and divided in three sections. On the top, the Main Menu is displayed. This menu is comprised from five main areas: Goals, Actions, Teams, Results, and Community. At the bottom of the page two columns are displayed. The first column contains the 'My Actions' area where the actions that are assigned to the 'logged in' user are displayed. The second column contains the

Announcements and the Events sub-modules, with the announcements and notices that have been published in the project.

The main software modules displayed in the Main Menu are divided into five key areas. The Goals area reflects the 'goals' of the innovation project. The Actions area reflects the 'actions' element of the project (tasks, deliverables, milestones). The Teams area on the other hand represents the resources of the innovation project. The Results area corresponds to the 'results' element of the innovation project, while the Community area reflects the 'community' element of the innovation project. These five areas effectively encompass the entire project innovation process and provide the team members and the project managers with tools necessary for its management. Each area has a number of modules that allow the project team members develop competence within specific areas. The software system eliminates the duplication of information as the data may be shared between different modules and sub-modules within the system and provides an environment for the effective collaboration between the virtual team members.

4.4. Goals

The Goals module of the menu corresponds to the Project Goals in the innovation project and allows the generation of four different forms relating to the project goals. These are Statements, Requirements, Objectives, and Indicators. They provide a structure for formulating the goals element of the project innovation process. Through the software the project management are encouraged to correlate the project's requirements, objectives and indicators and formulate a project's vision statement.

4.4.1 Requirements

The Requirements specification represents the internal and external demands on the project from key stakeholders. It involves identifying the stakeholders, defining the stakeholder

requirements and defining all change ideas. Each requirement has an acronym and a title. The title of the requirement has to be short and to summarise in few words the quintessence of the requirement, a more elaborate explanation can be given in the description field. Each requirement corresponds to one of the demands of one of the project stakeholders. The stakeholder can be selected from a drop-down list of stakeholders that includes the project funding agency, the project partners and team members, etc.

4.4.2 Indicators

The performance indicators forms that can be created in the Indicators sub-module in Goals area are metrics used to measure the performances of the project. They permit the continuous evaluation and improvement of the project performance. The Indicators need to be concise, easily measured, time bound and realistic. The Firm Target field indicates a short-term goal for a specific indicator. Stretch Target indicates a goal for up to the end of the project and its value is more a ‘best performance’ point than real allowing team members to 'stretch' their efforts to achieve them.

The four forms of Statements, Requirements, Objectives and Indicators permit the definition of goals that the project team will pursue, and their communication to the team members through the various views of data. The Goals module provides the team members with various views of data entered through other modules. The Deployment sub-module includes views that identify the relationships that exist between requirements, objectives, indicators and workpackages, and tasks.

4.5. Actions

The Actions module represents the actions in the innovation project. These actions are carried out to successfully achieve the project’s goals. The software sub-modules of Workpackages, Tasks, Deliverables and Milestones are the project’s activities.

The Tasks sub-module is designed to structure and capture the minimum information necessary to enable the team members to implement the action and monitor the evolution and the success of the action's implementation. When the user clicks on the Tasks hyperlink in the CPIM home page (see Figure 5) a page with all the tasks is displayed as pictured in Figure 6.

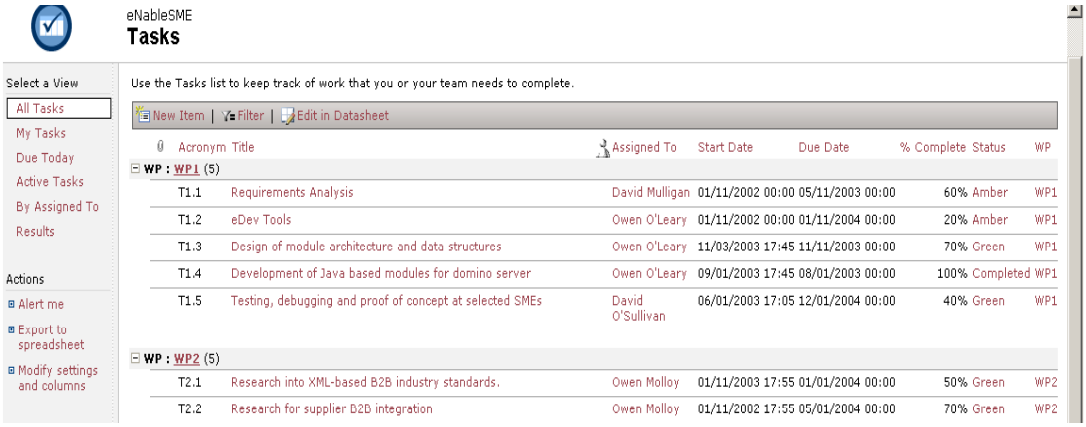


***Figure 6**: View page for Tasks module- Screen Shot*

In this view all the tasks in the project are listed. The main information about each task is displayed. The information in a list can be grouped, sorted or filtered. New columns can be added to the view and new views can be created. The team members can also subscribe to the page to be alerted every time the information is changed. The team members can drill down into each task record by clicking on the name of the task. A page with the task's details will be displayed. The team members can add, edit, view subscribe and delete one element from the list. The web pages corresponding to these functions are implicitly generated when the module is created. Figure 7 depicts the web pages corresponding to these functions for the Tasks module.

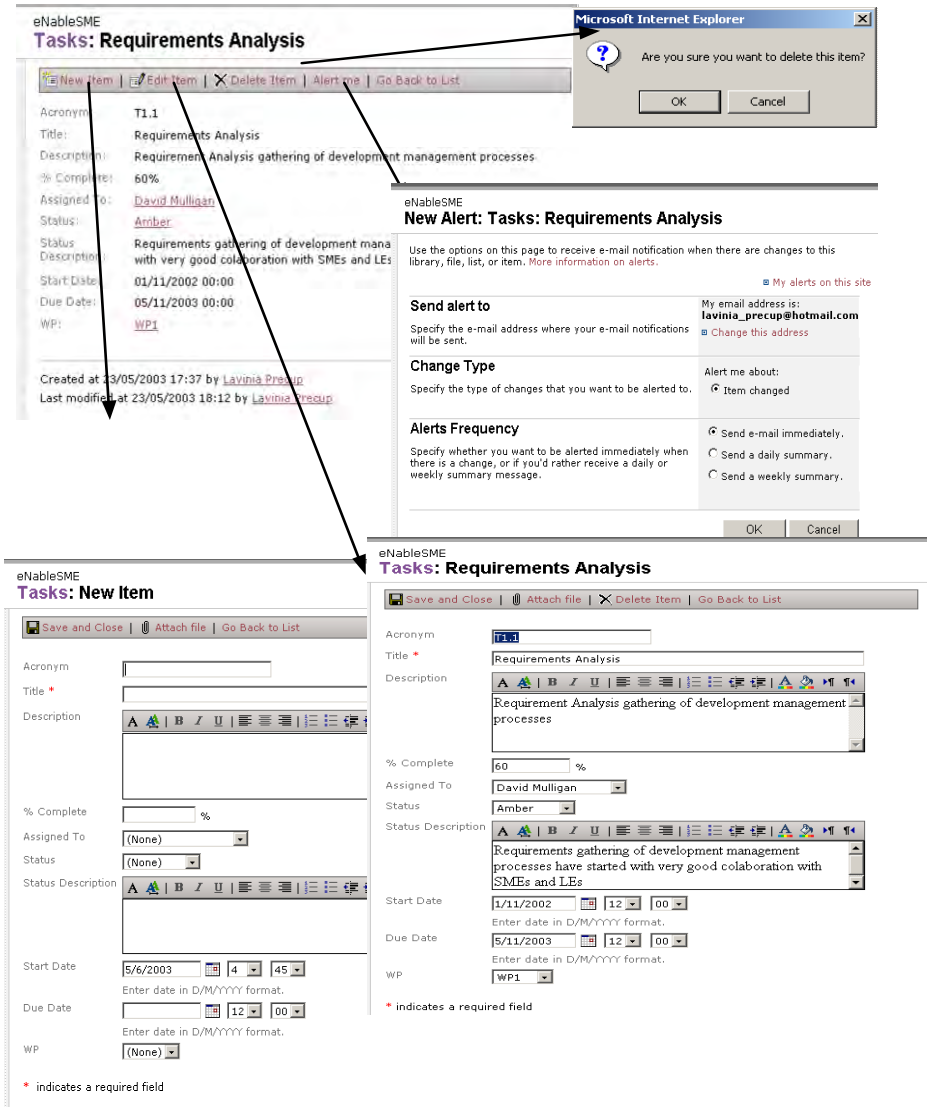


Figure 7: the functions for the Tasks module – Screen Shot

Figure 7 depicts the structure of the tasks forms. The Priority field is used to indicate the importance of the task for the success of the project. The %Complete and Status fields are used to monitor the evolution of the task implementation. The value of the Status field can vary from Red (the task requires immediate attention), to Amber (there are some minor problems in implementing the task), to Green (the task implementation is going well) to Complete (the task has been implemented). The Status Description field gives more details

on the status of the task. Each task is part of a WorkPackage and has an acronym, title, a start and an end date, and a responsible that has to ensure that the action is well implemented and it meets the project goals. The tasks can be tracked back to the project goals using the data from the Requirements field. Attachments can appended to each record in the list. The feature allows the team members to attach relevant documents and files to the tasks. For example if there are few individuals that work on the task they can attach the necessary files to the task. This way all the information needed for the correct implementation of the task is up to date and is held in one single place.

4.6. Teams

The project constraints also influence the implementation of innovative actions in the innovation project and are reflected through Teams module in the first page. The Team module addresses the constraining element of the innovation process. This module allows the identification of the human elements that interact within the innovation process and permits the effective management of their involvement. The Team module manages the human resources of the project through the two sub-modules: Partners and Individual. The Partners sub-module holds all the information about the organisations that are partners in the project. The Individual sub-module allows the capture of information about the project team members, their name, personal details, biography, and role in the project etc. Each Individual in the project has a user name assigned by the administrator of the CPIM, works in an organisation and has a status (role) in the project, which can be Research Engineer or Project Manager, etc. Other information that is captured about the individuals involved in the project is: the email address, phone, extension, web site address, and biography.

The virtual environment developed with Microsoft Sharepoint Services help allows the assignment of human resources to the development and implementations of innovation actions. The software functionality permits each team member to visualise the actions that

are assigned to him/her (Figure 5). The Team module acts as a communication tool, providing the project team members with up to date information about the other individuals and organisations involved in the project.

4.7. Results

Actions are implemented with the purpose of contributing to the project goals achievement. The Results module allows the project management to evaluate how successful the implementation of the project is. The goal-results relationship allows the project managers to evaluate the performance of the project and take corrective actions so that the project goals are successfully implemented. The results information is available through the WorkPackages, Tasks, Deliverables, Milestones, Indicators, and Objective result views, to all the team members so that they can act in order to correct the problems that appear. This gives them empowerment, helps the knowledge generation of the team members' knowledge, and reduces their resistance to the changes adopted by the project managers.

The Exception sub-module identifies all the actions and goals that require immediate attention due to their red status. A red status signifies that there are problems with implementation of the action or project goal achievement. Action needs to be taken to remedy the evolution of the action implementation so that the status of the actions and/or goals improves.

4.8. Community

The Community module offers the team members tools that facilitate and sustain the collaboration between the team members. There are three types of collaboration that are ensured through this module. These are asynchronous collaboration, distributed synchronous collaboration, and distributed asynchronous collaboration. The fourth type of collaboration, face-to-face collaboration, very important in sharing tacit knowledge is

ensured as well trough meetings and workshops. The tools available for sustaining and facilitating the collaboration between team members are: Documents, eDemos, Publications, Meetings, Announcements, Discussions, Templates, and Links.

Documents sub-module allows the users to share documents, edit, discuss over the changes and subscribe to the documents so that they will receive an email when a change appears in the library or document they subscribed to. Versioning control, access control and send it for review is also available.

The Publications, eDemos, Links, and Templates sub-modules allow the dissemination of knowledge between the team members in different forms (documents, links to outside sources of knowledge, presentations, etc.).

5. Discussion

Microsoft Sharepoint Services (Sharepoint II) provides a platform that facilitates an easy implementation of a virtual team environment. The platform comes with a friendly user interface and provides features that ensure an effective and successful collaboration and knowledge sharing between the team members. Microsoft Sharepoint Services fulfils the main requirements for the successful implementation of a virtual team environment.

The importance of the collaboration has been underline previously in this paper for the successful implementation of a project that involves virtual teams. Sharepoint Services supports three of the four types of collaboration identified by Anuba et al. (2001), which are implemented in the case study virtual environment. These types are:

- Asynchronous collaboration: that is supported by the Announcements and Notices modules, which display information in the form of notice boards and also by the Documents module that allows versioning and discussion over documents.
- Distributed Synchronous Collaboration supported by the Discussions module and by the integration with MSN for on line chats
- Distributed Asynchronous Collaboration supported through the integration with Microsoft Outlook which allows the users to subscribe to be alerted whenever the information on the site is modified.

Lipnack and Stamps (Lipnack and Stamps, 1997) and Pellecchia (1998) have identified the main requirements that have to be fulfilled so that the virtual team can successfully work and fulfil its functions. This requirements presented in the Virtual Team section of the paper are also fulfilled by Microsoft Sharepoint Services:

- Collaborative efforts are supported by Sharepoint which allows co-authoring, document version controlling, group editing and reviewing, annotating, sharing information and ideas, approval and author tracking for the documents.
- Communication is supported through the three types of collaboration implemented in Sharepoint.
- Co-ordination of the virtual team is also supported as all the teams actions are tracked. The evolution of work in the project is monitored in the case study environment in the modules of the Results area in the first page.
- Information sharing through document sharing and discussion forums, chats, links surveys, notices and announcements.
- Managing and tracking, project management are also supported by the Microsoft Sharepoint.

- Analysis and decision making tools to support collaborative examination of information, consideration of possible decisions, and development of best decision are the surveys and issues modules implicitly provided by Sharepoint and in the case study the Exception module which lists all the project actions that are not meeting their targets, fail to implement the project goals and require immediate attention.
- Microsoft Sharepoint encourages the development of collaboration between people who work at a distance from each other and who might not be from the same culture is a new organisation challenge through its three types of collaboration modules and through the integration with Microsoft Outlook that allows the translation of documents from one language to others.

Project innovation cannot be successful unless the knowledge and information in the project are effectively captured, shared and internalised by the project virtual team members. It is important for the success of the project that the new issues that arise from the virtual team management, and collaboration are addressed as the effective management of them will lead to the achievement of the project goals, and formation of knowledge that will not be lost at the end of the project but will be used to successfully implement new projects and continue to innovate. The virtual team environment developed using Microsoft Sharepoint Services effectively addresses this issues and allows for an easy and user-friendly management of the innovation project and of the virtual team working on the innovation project. At the time of going to press, Sharepoint II is in beta test mode and due for release in the Spring of 2004.

First Monday, Volume 16, Number 5 - 2 May 2011

PEER-REVIEWED JOURNAL ON THE INTERNET

Digging into data using new collaborative infrastructures supporting humanities-based computer science research

by Michael Simeone, Jennifer Guiliano, Rob Kooper, and Peter Bajcsy

Abstract

This paper explores infrastructure supporting humanities–computer science research in large–scale image data by asking: Why is collaboration a requirement for work within digital humanities projects? What is required for fruitful interdisciplinary collaboration? What are the technical and intellectual approaches to constructing such an infrastructure? What are the challenges associated with digital humanities collaborative work? We reveal that digital humanities collaboration requires the creation and deployment of tools for sharing that function to improve collaboration involving large–scale data repository analysis among multiple sites, academic disciplines, and participants through data sharing, software sharing, and knowledge sharing practices.

Contents

1. Introduction
2. The research driving the collaboration: Humanities-based computer science research
3. The logisitics of cooperation: Collaborative infrastructure to address specific challenges of collaboration
4. Summary

1. Introduction

At the core of humanistic research is the search for primary source information. Whether these sources include manuscripts, material culture objects, maps, or other forms like video and digital media, successfully navigating corpora of material has long been the substance of valuable research contributions. While scholars have traditionally worked by personally extracting data from archives and forming their own collections of information surrounding thematic or historical periods, the advent of the digital archive has posed its own unique challenges. Although they are promising in the resources they make available as virtual content, digital image archives are also potentially daunting. Digital archives can house so much data that it becomes impossible for scholars to evaluate the archive manually, and organizing such data becomes a paramount challenge. While individual scholars often commit the hours of labor needed to visually examine hundreds of images, the enormous size of some archives precludes the ability of an individual or even a team of individuals to examine each data element. Additionally, the kinds of questions that can be asked and answered are constrained when the research data is not well organized and synthesized.

With computational tools, digital archives can reveal more than they obscure by providing organizational frameworks and tools for analysis. However, these tools — in the guise of metadata organization, indexing, searching, and analytics — are not self–generated. They

require the combined work of humanists with their interdisciplinary questions and computer scientists with their disciplinary approaches to partner with one another to produce viable research methodologies and pedagogies. As interdisciplinary collaborations are becoming more common, aligning the interests of computer scientists and humanities scholars requires the formulation of a collaborative infrastructure for research where the approaches, methodologies, pedagogies, and intellectual innovations merge. While the concept of shared resources in a "cloud" is gaining popularity for office and document–based collaborations, we maintain that constructing a "cloud" of collective resources for use in researching large image archives across multiple disciplines and institutions requires a specific design that accommodates the communicative demands of multi–disciplinary academic research, the intricacies of intellectual property and publication, the needs of software developers working from remote sites, and the difficulties of serving large amounts of data for collective examination. In other words, the model we have developed tightly integrates practices and technologies for data sharing, software sharing, and knowledge sharing/communication.

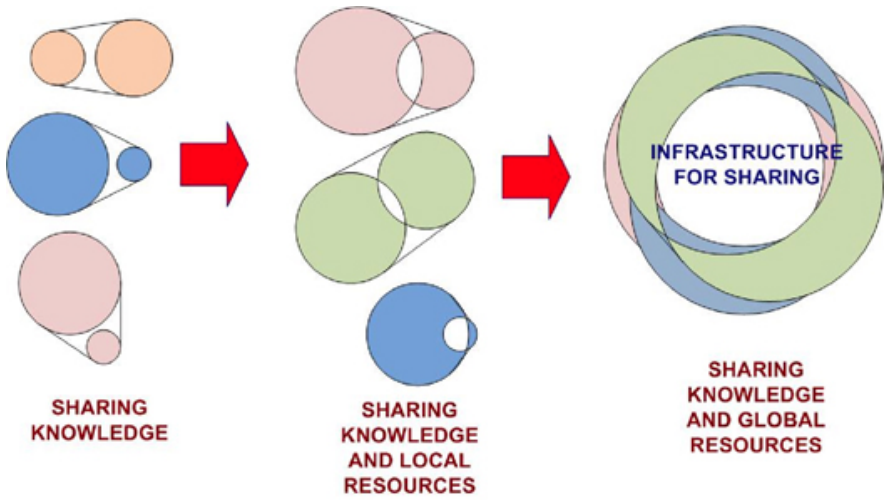


Figure 1: Paradigms for computer science and humanities collaborations. Circles of different sizes represent research and resources of each collaborator with lines connecting them to represent project collaboration. The scenario on the left indicates the most basic kind of sharing, the scenario in the middle shows a series of partnerships with some shared resources, while the scenario on the right demonstrates a shared and generative infrastructure for circulating data, knowledge, and hardware among multiple collaborators and institutions.

This article explores the construction of one such infrastructure for sharing in digital humanities and computer science collaborations. We collectively ask: Why is collaboration a requirement for work within digital humanities projects? What is required for fruitful interdisciplinary collaboration? What are the technical and intellectual approaches to constructing such an infrastructure? What are the challenges associated with digital humanities collaborative work? In doing so, we reveal that digital humanities collaboration requires the creation and deployment of tools for sharing that function to improve collaboration involving large–scale data repository analysis among multiple sites, academic disciplines, and participants through data sharing, software sharing, and knowledge sharing practices. Increasing the number of participants from diverse disciplinary fields makes the logistics for that collaboration of crucial importance. The suite of tools developed and employed under the Digging into Image Data to Answer Authorship–Related Questions (DID–ARQ) Project to manage this communication and collaboration and help reduce the start–up and time and costs is instructive for future large–scale, multi–institutional projects. It is our goal to not only demonstrate the challenges and solutions that emerged throughout our work, but also to provide a template for future collaborations that involve multiple datasets in geographically distributed locations.

As a case study, we use our work in DID–ARQ awarded as part of the Digging into Data Challenge Competition (DID Challenge Web site). Funded by the National Science Foundation (NSF) and National Endowment for the Humanities (NEH) from the United States, the Joint Information Systems Committee (JISC) from the United Kingdom and the Social Sciences and

Humanities Research Council (SSHRC) from Canada, DID-ARQ consists of an international, multi–disciplinary team of researchers from the University of Illinois (U.S.), National Center for Supercomputing Applications (U.S.), Michigan State University (U.S.), and the University of Sheffield (U.K.). The DID–ARQ team is conducting work to formulate and address the problem of identifying the salient characteristics of artists from two–dimensional (2D) images of historical artifacts. Given a set of 2D images of historical artifacts with identified and unidentified authors, our project teams aim to discover what salient characteristics make an artist different from others, and then to enable statistical learning about individual and collective authorship.

Our collaboration infrastructure produces an innovative research environment because of its ability to integrate multiple collaborators (knowledge), content types (data), and research tools (software and hardware) across multiple locations. Figure 2 shows an example of an infrastructure for sharing that was deployed for our DID–ARQ Project. The three sites involved used a cloud of technologies for sharing images, software, hardware, and knowledge with the sharing policies defined by the memorandum of understanding. Because we employed these enhanced tools for sharing, our collaboration has produced a uniquely comprehensive exchange between computing and humanities stakeholders, one that lets individual researchers or teams of researchers dynamically form their research goals the better to cooperate with other researchers within the Project. Indeed, these tools have enabled, archived, and published every phase of this multi–campus, multi–archive, and multi–disciplinary endeavor. This infrastructure allows us to add additional sites in the future that can connect themselves to this cloud, become familiar with previous collaboration work, see the data currently available, and quickly get integrated into the whole project. Providing this structure allows new partners quickly to start adding to the cloud of technologies thereby increasing the speed of deployment.

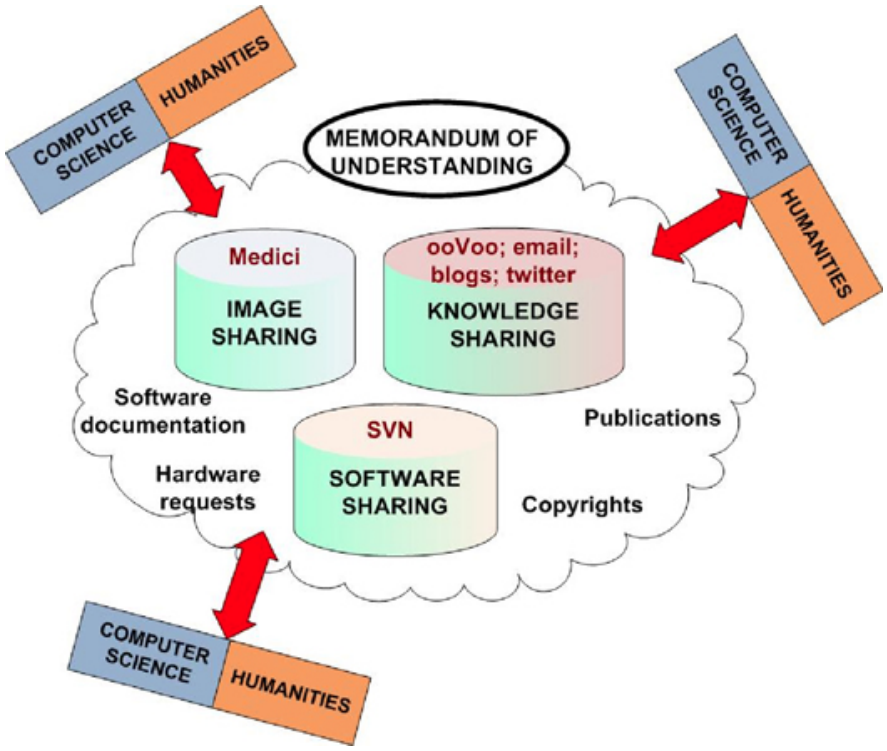


Figure 2: An example of an infrastructure for sharing that was deployed for the DID–ARQ Project. An example of an infrastructure for sharing that was deployed for the DID–ARQ Project. The three sites used a cloud of technologies for sharing images, software, hardware, and knowledge with the sharing policies defined by the memorandum of understanding.

2. The research driving the collaboration: Humanities-based computer science research

Each of the individual projects of the DID–ARQ collaboration starts with a common research question: how does one examine an archive of images when that archive is too large to examine manually? In framing that question through collaborations with geographically remote scholars, however, a second important question arises: how to share these archives and communicate their salient characteristics with research partners? Historical maps of the Great Lakes region, The Quilt Index, and Froissart Manuscripts all represent hundreds to thousands of digitized images, and every digitized item, no matter its historical context or intended purpose, is a potentially rich store of information for scholars of history, literature, and culture. Because of the sheer volume of labor required for research across the archives in question, "when" the researcher might complete his or her study of the entire corpus becomes a question of "if" that research is even possible. In this section, we will briefly explore the humanities questions from each of our humanities partners and then examine how remote collaborations have affected the questions being posed.

2.1. Pattern driven classification of quilts

The Quilt Index is a growing research and reference tool designed to provide unprecedented access to information and images about quilts held in private and public hands. A joint project of The Alliance for American Quilts, MATRIX: Center for Humane Arts, Letters and Social Sciences Online at Michigan State University and the Michigan State University Museum, the Quilt Index provides an opportunity for in–depth exploration of quilts through the lens of thousands of digital quilt images and text drawn from public and private collections. Two main issues, the sheer volume of data and variation in the text documentation, limit the capacity for any individual to examine the entire corpus of the Quilt Index thoroughly. The variation in the quality and format of the documentations lends itself to challenges of organization. While each quilt has a set of descriptive fields that contain necessary information and controlled vocabularies specific to quilt research and study, for each quilt the original textual documentation must be matched to particular descriptive fields. While differences in existing text strings can be accounted for through the search programming, entailing, for example, algorithms for effecting the simplest reconciliation of spacing and capitalization, or more complicated cross–references for known spelling variation and alternate published pattern names [1], cross–reference lexicons are not effective to account fully for unique names or to cover every local variation. Moreover, in some cases key information may be unknown or even intentionally left out during the documentation process, such as a quiltmaker's affiliation to an organization like the Ku Klux Klan. These challenges when parsing the text documentation create intensive labor for processing quilts. The challenge as Gregory Crane famously asked about books, is what can a humanities researcher do with a million quilts? (Crane, 2006) In the first place, the physical collections of quilts are geographically distributed, which rules out most side–by–side comparison. Moreover, even if institutions allowed physical access to the collections, a scholar could not possibly unfold and physically look at more than a few quilts at a time and then only at an oblique angle, as most quilts are the size of double beds. Even in a digital environment designed for users to frontally see and compare images, close analysis of every digital quilt image is time–prohibitive when the number of digital images reaches into the tens of thousands.

2.2. Knowledge of cartographers of early French and British maps

Early mapping of the North American continent (1640–1820) generated sizable archive of ornate atlases and sheet maps. Studying seventeenth and eighteenth century maps in the context of modern understandings of cartography reveals distortions that were the inevitable results of comparatively primitive technologies, difficulties of travel by sea, rivers, and overland, and inaccurate and conflicting data from explorers and traders. Early mapmakers relied on various methods of recording distance that recorded other information as well, as techniques ranged from dropping a log overboard and watching its movement, to examining the eclipses of Jupiter's moons. Changes in coastal currents, however, could warp the apparent distance traveled by passing ships, cloud cover could block out astronomical reference point, and geographical features like bays, rivers, and mountain passes were both consciously and unconsciously exaggerated to appear more appealing to shipping and commercial interests. Furthermore, European sailors could not calculate longitude until the later part of the eighteenth century, and before then, climate affected primitive timepieces and thus any measure of east–west distance.

Yet if plotting the regions around the Great Lakes without error was impossible until the later

development of more accurate compasses, timepieces, and recording equipment, early maps still contain a vast amount of valuable data. There is a rich history to what seem now to be mere mistakes and miscalculations — a history that discloses both the development of technologies for measuring distances and the ways in which French and English mapmakers interpreted the geographical data that they had at their disposal.

In the study of maps, evaluating each map's depiction of geographic regions and waterways, from their relative size to their purported location on the Earth, increases in difficulty as the maps become more intricate and detailed. Given that there are hundreds of available maps of North America produced by European and Native mapmakers, hand examination, though useful in limited qualitative studies, makes for a nearly prohibitive workload if scholars wish to pose any research questions across not only 40 maps at a time, but larger portions of the archive as well. This archive contains valuable data needed to address fundamental research problems in colonial history, the history of science and technology, historical climatology, and the history of print culture.

2.3. Artistic and scribal hands contributing to manuscripts

The Froissart Manuscripts corpus consists of more than 6,000 high–resolution image files captured photographically from 12 manuscript volumes (~2TB of data). These manuscripts were all produced between ca. 1408 and ca. 1418 under the supervision of Parisian bookseller Pierre de Liffol, who seems to have glimpsed a market opportunity for the production of luxury copies of the Middle French Chronicles of John Froissart (ca. 1337–ca. 1404), now a key source for the study of the Hundred Years' War between France and England. De Liffol produced these volumes for clients in the service of king Charles VI of France, though the illustrations to at least one of them testify to a client with pro–English sympathies. The text of the Chronicles was copied from exemplars by at least two teams of scribes. The miniatures were entrusted to two artists' workshops: those of the Giac and Boethius Masters. Secondary decoration (the initial letters and borders) may have been painted by still other artists. Once all of the copying and illustrating was done, the quires were reassembled, sewn together and provided with oaken boards covered with leather or velvet.

The virtual manuscripts allow the scholar to consult the content at their leisure and manually compare across different collections. Furthermore, the digital surrogates offer the opportunity to perform more complex analytical comparisons across the expanding collection than a human is capable of undertaking. By considering the detail and aspect of each of the manuscript folios, similar scribal hands and artistic objects can be clustered together for scholarly comment and thus help construct a pattern of work across them with the aim of determining authorship and work patterns. The algorithms will also help scholars to adduce more discriminating features than hitherto arrived at for characterising the distinctive handiwork of the Giac and Boethius Masters as well as reveal information about the hands responsible for the primary and secondary decoration of these manuscripts. More specifically, can we discern the presence behind the labels 'Giac Master' and 'Boethius Master' of more than one individual (*e.g.*, a group of artists working in a common style)?

2.4. Need for a new collaborative infrastructure

Scholars in the humanities, including those above, are faced with a set of specific research questions that pertain to their individual disciplines and an archive that tests the limitations of their disciplinary methodology in scope and size. For each humanities project, the questions of managing and sharing data across these repositories become a key challenge. Each project has its own metadata, organizational structure, and image repository. Although each project focuses on its particular issue, yet, the need to explore authorship unites them all. Furthermore, it is not only the authorship question that unites the smaller projects but also the software that can be reused. For example, as illustrated in Figure 3, the same software algorithm has been used to find objects of interest in Froissart's manuscripts (*e.g.*, armor, faces) and lakes in historical maps. Thus, sharing data and software with hardware is critical for resource limited research in humanities.

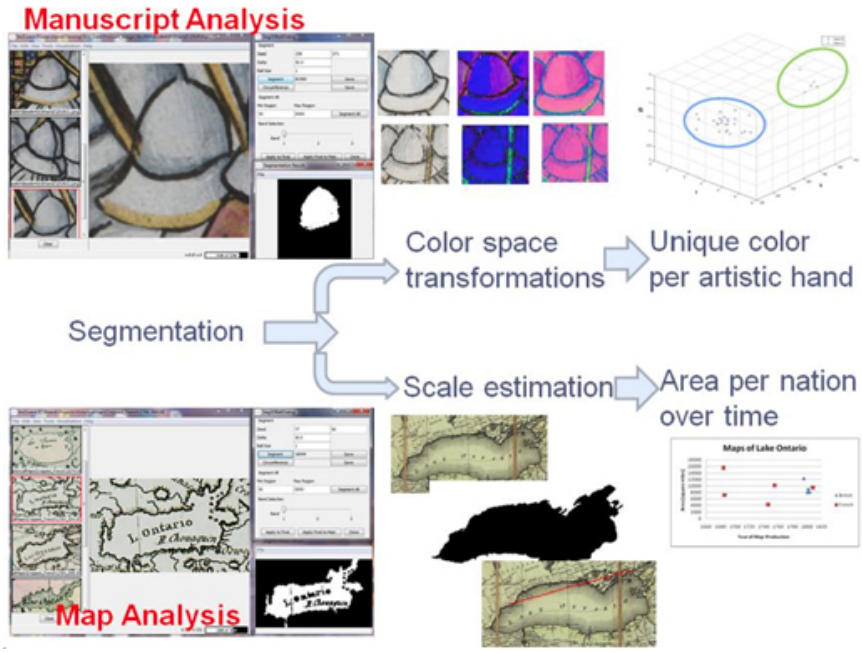


Figure 3: Illustration of the need to share software that can perform model–based segmentation of images. The same segmentation algorithm is applied to finding armor in Froissart's manuscripts and lakes in historical maps.

The challenges faced by both the individual projects and the overall DID–ARQ group for collaborating among those individual projects necessitated a new collaborative infrastructure. Nonetheless, the problem of designing such a resource poses several questions. With such large–scale archives, how is it possible to share information effectively and purposefully with project partners? If hundreds of images press the capabilities of manual examination, how are other contributors able to craft interpretations of the respective corpuses? How can a computational scientist design a solution to an image–processing problem for a group of images that is this vast and also geographically remote? As we initiated our collaboratory, these questions began to structure our enterprise: how could scholars at three different sites working with collections of visually disparate materials and with three disparate methods of image storage and management work together to reveal salient characteristics making one artist different from another regardless of the source material?

3. The logisitics of cooperation: Collaborative infrastructure to address specific challenges of collaboration

As outlined above, in formulating a collaborative humanistic project, we had to recognize the individual questions and the methodological and theoretical approaches of each humanities scholar. Yet, a single humanist exploring an image archives on their own is not the primary audience for this project. Rather, it is the intersection of teams of scholars all unified in their technical challenges of collaboration that drives this project forward. In this section, we outline how financial, legal, and research considerations resulted in the construction of a collaborative infrastructure that not only facilitated the interchange between scholars and scientists but also enriched the very nature of information exchange. With challenges ranging from country–specific and institution–specific policies for collaborations to technical problems related to sharing data, software and computational resources, and building communication channels for knowledge sharing and project coordination, we begin by noting an underlying issue with the framing of all Digging into Data projects, including the DID–AQR Project.

3.1. DID institutions and funding

Announced in January of 2009, the Digging into Data grant competition was sponsored by the NEH, NSF, JISC, and SSRC. In releasing the request for proposals (RFP), these organizations created one main RFP with addenda for each of the sponsoring organizations. Each institution within a DID team was required to field an international partnership involving at least two of the three participating countries: Canada, England or Wales in the United Kingdom, and the United States. While our complete team would create a single application that would be sent to the agencies, each individual institution was required to create their own budget and follow the rules of their assigned funder. In the case of DID–ARQ, UIUC/NCSA applied to NSF, Michigan State to NEH, and the University of Sheffield to JISC with each team completing its own budget and sections of materials with consultations from their targeted partner institutions. A single named primary investigator was charged with completing the application and submission process, yet each institution needed its own primary investigator to complete their local grant process and deal with their assigned sponsoring agency. In effect, although the project intent was to promote international collaboration through collaborative research, the application structure immediately necessitated sub–division through the required one–to–one relationship with each sponsoring agency and the adherence to the grant submission and management systems of each applicant organization.

This format extended throughout the funding mechanisms associated with the Project. Each institution created its own budget and, when awarded, received its own contract from its assigned sponsoring agency with no mechanism for sharing funds or for collaborative oversight. DID awardees had to design their own mechanisms to leverage geographically distributed expertise and resources, as well as independent sources of funding. The DID awardees did not need to outline an in–depth plan for how they would collaborate, just that they would collaborate internationally and in some discrete way. This underlying assumption of having all collaboration infrastructure and methodology in place at all sites resulted in proposals that largely ignored the challenges, approaches, methods and technologies needed to address a distributed collaboration, as this component of the proposal was not critical to obtain the funding. As a consequence, our awarded DID–ARQ Project immediately faced challenges of how to structure and fund the expense of the collaborative activities that were not institutionally based. This included start–up costs associated with the infrastructure technologies for sharing, and with the preparation of agreements governing the collaboration. For example, in our proposal for DID–ARQ, we noted that we would utilize AccessGrid, a peer–to–peer technology for large–format information sharing and videoconferencing. However, the costs for AccessGrid (US$136 per hour) were not figured into the initial budgets. With two and a half to three hours of technical meetings a month, costs for AccessGrid would exceed US$6,120 for the Project. With the lack of budgeting, these costs would have reduced available funds for other portions of the Project. Due to the lack of a formal mechanism for parsing start–up costs between teams or a shared pool of funds, these types of costs would have been a burden on one institution rather than on the team as a whole. As a result, the assigning of start–up costs became a *de facto* decision of convenience with the institution shepherding particular aspects of the collaboration bearing those costs.

3.2. Legal and ethical aspects of scholarly collaborations

A noteworthy issue in initiating the DID–ARQ Project involved navigating various policies specific to countries and institutes that apply to collaborations and sharing. Before project work could begin, the DID–ARQ team had to negotiate the legal aspects of collaborations regarding the conduct of the Project, publicity, publications, copyright, intellectual properties and liabilities. In accommodating U.K. and U.S. government–specific policies, as well as the institutional policies of the University of Sheffield, MSU and UIUC, it was readily apparent that formal agreements would have to be set in place. In part, this was necessitated by team concerns about project liability associated with data access. Images associated with the projects could be lifted from project Web sites and publications and used in violation of copyright laws. As such, while work on the project began while the agreement was being negotiated, all due consideration was given to limiting access to data and images by the general public. Initially drafted by the University of Sheffield, and reviewed by MSU and UIUC Grants and Contracts offices, the memorandum of understanding associated with DID–ARQ documents how all project principal investigators, their graduate students, postdoctoral scholars, and faculty staff will follow a prescribed set of behaviors in documenting the work of the Project, the publications associated with Project results, and intellectual property contributed by project participants.

One important issue for the institutions involved was ascribing ownership of images to their respective archives. The DID–ARQ team discussed extensively the rules for publishing materials with the images that would be shared across all sites for the software development

and testing purposes. The digitization of Froissart's manuscripts was paid for by several private donors and by sponsors of prior projects in the U.K. [2]. The acquisition of quilt photographs was subsidized by MSU and by funding from a variety of public and private sources [3] just as the digitization of historical maps from UIUC was sponsored by the University of Illinois Libraries. The data stewards on the DID project had to decide how to select images for sharing and how to formulate the acknowledgement if an image was included and therefore referenced in a publication or presentation. As a consequence, the collaborative technology for image sharing had to provide a mechanism for entering author, copyright information, credit, and acknowledgement information that would stay affiliated with an image including any derived images or datasets. Agreement on the legal aspects of this collaborative project would have been much more difficult if the partner institutions had not all been educational ones. While working with private collectors, companies or philanthropists might bring new images to the research arena, the objectives of partners and the legal implications of different objectives might require devoting more time to this legal aspect than was the case with our effort. Construction, negotiation, and signing of the DID–ARQ MOU was completed in roughly three months as each institution was able to utilize its local legal and contract staff in creating the agreement. Significantly, institutions should plan at least four–five months to take care of the legal aspects of collaborations particularly when dealing with non–educational partners.

We organized the legalities of this effort, in part, by our observation that the grantors impose different criteria on measuring success of a project with respect to the expected project outcomes, and reporting methods. The Joint Information Systems Committee (JISC) outlined an expectation for more applied research than did the National Science Foundation (NSF). Given the unwritten rules about sponsors' evaluations — evaluations driven by the reviewers coming from the communities being supported by these sponsors — we ascertained that the National Endowment for the Humanities (NEH) desired an emphasis on academic monographs about discoveries of new facts from historical content while NSF would rank higher journal papers about the computational and statistical methods needed to obtain those discoveries. The differing interests of each funding agency underscore not only the difficulty of bringing together multiple funding agencies in providing funding but also the difficulty in meeting the expectations by the proposers and awardees. While individual team members clearly understood these differences, they had to be taken into account when legal agreements were drafted and the division of work outlined. Graduate students focused more on basic research (*e.g.*, how to design computer algorithms to automate authorship discovery) while full–time staff pursued applied research (*e.g.*, how to support tagging and commenting when images were shared across sites). Expanding to publications, the team agreed that computer science journal papers can use images to demonstrate the algorithmic performance and computational scalability with appropriate acknowledgments, while books written by humanists about the interpretation of historical content can use computer–generated results and references to the images used with appropriate acknowledgments.

Another important question is that of authorship of articles, papers, journals, and books written by members of the DID–ARQ Project team. In an academic climate where "publish or perish" is still a measure of success for both humanists and scientists, determining the intellectual ownership of research publications has challenged our own parameters of authorship. The awarded DID–ARQ proposal included 18 participants, and the current project involves an additional four students and three full–time engineers. Multiple authors is more common in the sciences than in the humanities, but the number of individuals contributing to this project is unusually large. A question arises about publishing an article with 25 co–authors or a subset of co–authors, or in other words, about individual and collective contributions. While we do not know a fair answer to this question, a policy about co–authorship has to be discussed among the team members. The DID–ARQ Project team proposed guidelines that acknowledged that significant intellectual contributions merited authorship and co–authorship credit; the contribution of content materials related to the individual projects drawn from the original proposal project materials without additional intellectual investment in the resultant publication did not merit co–authorship. In effect, each scholar and scientist could receive acknowledgement of their content contributions (generally framing questions or summaries of work undertaken), but this acknowledgment stands apart from authorship credit. Authorship credit comes only with significant investment in the writing and constructing of interpretations and arguments in new publications.

Generally, while a member or members of the team initiate(s) a publication or public presentation, DID–ARQ relies on notification of the initiation of a resultant publication/presentation and the sharing of all works in progress in order to give all members of the team the opportunity to contribute. In this manner, DID–ARQ ensures an equal–work, equal–credit scenario where all members of the team, be they graduate students, staff or PI, can take ownership of the publication. This structural element ensures that all legalities of intellectual property and publication are met while still recognizing the need for individual

autonomy by the teams making up the full collaboration. A variety of public materials have been generated through this apparatus. There has been a paper authored by a subset of team members on preliminary materials that became part of the DID–ARQ partnership, three presentations on two different continents that were authored singly but include DID–ARQ acknowledgements, and two publications in process that include acknowledgements, co–authors, and authors of differing degrees.

3.3. Data sharing

Given the challenges posed to our research and collaboration by the kind and scale of our image data, it was crucial for our DID–ARQ team to establish a common means to collect, share, annotate, and examine large amounts of image data. To address these challenges, we utilized a previously developed prototype of a multimedia content repository called Medici [4]. Medici is unique as a shared project workspace: it can accommodate the upload of all file formats in a common environment with evolving rendering and capabilities based on an expanding list of project requirements.

From a number of possible choices for applications supporting cloud computing and file sharing, the DID–ARQ team selected Medici because it allowed for distributed access and management of large-scale private image collections. Commercial systems such as Flickr (Flickr, 2010), Panoramio (Panoramio, 2010), or TinyPic® (TinyPic®, 2010) were not suitable because they contain no ways to protect copyrighted content from copying or download, a crucial feature given the constraints imposed by each individual image archive. In comparison with open source distributed data management solutions such as Storage Resource Broker (SRB) or iRODS (iRODS, 2010), Medici provided extensive visualization capabilities for viewing multimedia content including very large size images which have not been available in iRODS but have been critical for working with DID–ARQ images. Finally, the social networking features present in many commercial solutions, such as personal tagging and commenting, were very helpful during data sharing but unavailable in the open source solutions. In the future, we foresee the need to add image–based search (see TinEye, 2010) in addition to text–based search.

In Medici's Web–based file browser, research papers and images appear side by side as thumbnails, and custom collections showcase images and documents that elucidate specific research problems. Furthermore, Medici lets its users examine shared files in their full resolution without downloading them. Built into the file viewing system is a preview window that uses Seadragon, a library built by Microsoft Labs that lets users smoothly zoom in and out when viewing large image files.

Thus, where sharing gigabytes of data was an obstacle on other platforms, it is now through Medici that there is an opportunity to provide a rich multimedia experience for collaborators. For example, large amounts of quilt images, instead of requiring individual project participants to commit significant time and resources to downloading and viewing the full–sized images, can be examined now in their full resolution through a Web browser. Unlike existing systems, including Flickr and YouTube, with some of the same functionalities as Medici, one of the draws of the Medici system is the ability to create a private data collection. The data that is stored in the system is only accessible by those people who have been given user access to the system and once the project is finished the data can be safely removed without a trace. For cross–university projects such as DID–ARQ, the ability to control access to data and the ability to remove it at the end is very important so that all legal requirements are fully met.

This ability to share image data at full resolution is crucial for our computational scientists and humanities scholars. Because these are photographs of physical artifacts like maps, quilts, and manuscripts, their study requires as much detail about their materiality as possible to be made available. The content is *digitized* not digital, and thus research on the digitized archives is not possible without maintaining fidelity as data moves from artifact to digital image to remote research location. For instance, scientists at Illinois cannot evaluate how to adapt the shape segmentation algorithms they have developed for the study of historical maps to help with the study of medieval manuscripts without a detailed replica of the object they will be analyzing, complete with the details, textures, and imperfections of the original. All collaborators were able to upload, annotate, and examine their own collections of images as developed, hosted, and deployed by NCSA/UIUC, and to do the same with collections built by other participants.

To use Medici, scholars upload files without having to specify much more than the file itself by using the standard file browser interface, or by clicking and dragging folders into a drop box. Single files are added to the default file collection where folders become a new collection of files based on the folder name. Once users add a file to a Medici repository, the system extracts basic file information such as the filename that becomes the title by default. The system assigns a unique URI, and it executes extraction or analysis services automatically.

Afterward, the user can include any statement about the data, such as adding a tag, leaving a comment, or adding a specific metadata field to help describe or categorize the digitized artifacts, such as publisher, mapping organization, or region of production. Ultimately, then, Medici furnishes collaborators not only with the means to share and easily access large amounts of image data, but also to showcase proposed logics of organizing and approaching the archives in question.

3.4. Software sharing

In addition to data sharing, the DID–ARQ team had to share pre–existing software and manage concurrent software contributions from the three institutions. The pre–existing software included not only the Content Management Repository called Medici but also the Im2Learn library of basic image processing and visualization algorithms that can be applied to various image analyses, the Versus (Versus, 2010) library for content–based image comparison, the Cyberintegrator (Cyberintegrator, 2010) workflow for managing computations on distributed computational resources. The analytical capabilities come from the Im2learn library that provides a plug–and–play interface for adding new algorithms and tools. Due to the fact that the authorship questions are frequently based on a comparison operation, we have designed additional application programming interfaces (API) called Versus which allows everyone to contribute with comparison methods. Once the algorithms for image analyses and comparisons have been developed, they can be integrated into workflows (a sequence of algorithmic operations to reach the analytical goal) in Cyberintegrator workflow environment. Cyberintegrator is a user friendly editor to several middleware software components that enable users to (1) easily include tools and data sets into a software/data unifying environment, (2) annotate data, tools and workflows with metadata, (3) visualize data and metadata, (4) share data and tools using local and remote context repository, (5) execute step–by–step workflows during scientific explorations, and, (6) gather provenance information about tool executions and data creations. Unlike other existing open source workflows used in eScience, such as Taverna (Taverna, 2010) or Kepler (Kepler, 2010), Cyberintegrator has a reconfigurable user interface by using the Eclipse Rich Client Platform (RCP), and captures all provenance information automatically using semantic representation (Resource Descriptor Framework representation). These characteristics allow a user to customize Cyberintegrator's interface to his/her preference and to mine the provenance information graph to reproduce trial–and–error research results or to gain knowledge about the use of data sets and software.

We have managed concurrent software contributions from the three institutions by setting up a subversion source control system [5]. This system allows for software sharing, remote access, software merging, version tracking and permission management.

3.5. Hardware sharing

The computational power of a single desktop has risen dramatically in the past few years. It is not uncommon to have a machine as desktop that has four computational units, each of which is many times more powerful than those found in desktop computers previously. The same growth seen in computational power can be seen in storage. Hard disk storage is now measured in terabytes instead of gigabytes. This growth in both computational power as well as storage has led scientists to tackle larger and larger problems in common computing environments. Still, running many processing steps (computational algorithms) on the same data set can take a few days. Similarly, running one fast processing step on a very large collection of data sets will also make the total time to process a few days or even months. Having access to large amounts of shared data and running the processing steps on the data will only make the computational time expand. Scientists purchasing modern desktops today expect to obtain the computational results faster than ever before. To match the expectations of scientists with the computational requirements, one needs not only high–performance computing hardware but also the software that can run in parallel and take advantage of the hardware.

High–performance computing hardware is becoming more mainstream in universities and companies. A cluster can be created using off–the–shelf personal computers that are connected using free software such a Linux and Eucalyptus, or by using solutions from commercial software vendors including Microsoft. If a cluster is not immediately available, the computation can be outsourced to places such as Amazon. Companies like Amazon sell computational resources to users for short periods, enabling them to process large amounts of data quickly. For the DID–ARQ Project, we have the advantage of having access to the high–performance computers at the National Center for Supercomputing Applications (NCSA). In addition, NCSA is currently in the process of creating the fastest public computer, while both MSU's High Performance Computing Center and Sheffield have smaller clusters that can be used as well. If any of the processing is projected to take a large amount of time, NCSA has offered to run the algorithms on its clusters assuming that the algorithms have been

parallelized.

3.6. Knowledge sharing and project coordination

With partners geographically distributed across three different time zones, verbal communication posed the first major challenge for knowledge sharing and project coordination. In our 18–month long collaborative, the University of Sheffield is six hours ahead of our western most partner, the University of Illinois (except for a few weeks when daylight savings did not match up). With bi–monthly teleconferences, finding a time that took this time difference into account and would accommodate all the partner institutions was paramount. In effect, Illinois was just beginning its workday as Sheffield was closing down its efforts. Further, many of the DID–ARQ team members were "targets in motion" simultaneously juggling other project schedules, travel opportunities, and professional contingencies that made knowledge sharing difficult.

At the core of the challenge of verbal communication were the disparate languages used when discussing DID–ARQ. The fourteen to 25 participating team members in the DID–ARQ Project are drawn from diverse backgrounds including computer science, history, English, art history, folklore, and interdisciplinary project management. Some members had computational backgrounds; others had little understanding of data management and organization. In the same way, while all team members were generally familiar with each other's materials due to previous collaborations, collaborators possessed little understanding of the complexities driving each of these humanities projects. Our challenge in communicating with one another was best illustrated by one of our initial issues in our grant application: the materials produced by our computer science colleagues referred to data while the materials produced by our humanities scholars referred to primary sources or source material. As we educated one another about the underlying meanings of data (which can be singular or plural and content–agnostic) and primary source (which is generally singular and always associated with a specific format), the realization was made that we would need to devote attention to defining not just these terms but also outlining the process by which we conducted our scholarship. Thus, it became essential to uncover not just the questions outlined above but also what the context was that lent these questions with reference to the issue of authorship. Highlighting the effort to uncover terminology and meaning was the negotiation of what constituted "progress" and "success" within the Project. The final deliverables for the project were outlined as (a) data about salient characteristics of an artist with respect to another artist and with respect to a group of artists, and (b) software for obtaining salient characteristics. Yet, as we began the Project, it was apparent that negotiating three differing projects would also necessitate explicating how each institution measured success. Was success participating in the required meetings? Meeting assigned deadlines for each person's individual task? Was it aiding one's collaborators in their tasks while simultaneously completing one's own? As this was experimental research, could "success" with the Project also be achieved by realizing avenues of research that did not yield concrete results?

Our solution to "targets in motion" and evaluating progress takes into account the need to document the project work and continually discuss our ongoing efforts. First, MSU initiated a DID–ARQ listserv that would be our primary method of internal asynchronous connection. All DID-ARQ members utilize that listserv for Project–related communication that related to the complete teams. Distribution of meeting agendas, papers/publications in progress, and informational dispatches are the primary contents of the listserv. That listserv served as an internal archive of written communications and allowed for observation by evaluators at MSU. Secondly, for public consumption, MSU authored a Web site outlining the project and its progress including a list of all presentations and publications (DID–ARQ, 2010). This site is available as a central aggregator for the project and can serve a mirror site for all Project institutions to disseminate for future presentations of the DID–ARQ Project. Accompanying the MSU sites is a DID–ARQ RSS subscription feed that delivers updates to interested parties as DID–ARQ progresses. Lastly, DID–ARQ established a video conferencing apparatus for meetings that would allow for remote participation from multi–country sites.

DID–ARQ staff quickly recognized the value of video conferencing software that could record audio and video conversations. This allows team members with scheduling conflicts to watch the meeting asynchronously and with the aid of the DID–ARQ listserv, contribute to the agenda either prior to or following the live discussions. Additionally, the "target in motion" nature of DID team members necessitated having software that could accommodate a minimum of four video streams: Sheffield, MSU, UIUC, and NCSA. The popular Skype, the free video conference technology of choice of most team members, could not accommodate either recording or multiparty calls at this point nor could the popular iChat. The communications solution was offered by ooVoo. Video conferencing software in the same vein as Skype, ooVoo accommodates up to six video streams, recording of audio and video, additional telephone

participants, and desktop sharing at a cost of US$440 per year. As outlined above, this is a significant cost reduction from AccessGrid which was initially considered by the Project. All DID–ARQ meetings are initiated by NCSA, who controls ooVoo access. NCSA logs in, thereby enabling the multi–party business features, and then calls the other DID–ARQ participants, who have downloaded the free, publicly available version of ooVoo. Using an agenda distributed through the DID–ARQ listserv, each call commences with the announcement that recording will begin. Following the meeting, these recordings are post–processed and uploaded to the Medici repository and tagged as technical meetings. This allows evaluators and participants to have a full catalogue of calls at their disposal to refer to at their leisure. Most significantly, this technology allows for continual re–consultation of discussions and agreements to confirm decisions and processes undertaken by DID–ARQ staff. It is an asynchronous evaluation tool for reviewers to establish the progress made bi–monthly. All aforementioned communication technologies provide a means for knowledge sharing and project coordination as illustrated in Figure 2.

4. Summary

The paper covers a range of research problems encountered in humanities endeavors that require computer science research and a collaborative infrastructure. We documented specific challenges as encountered in the Digging into Data to Answer Authorship–Related Questions project and described collaborative technologies deployed for the DID project. As the title indicates, there has been a growing number of Digging into Data problems in humanities that need to bring together geographically distributed expertise and resources, and therefore need new collaborative infrastructures supporting humanities–based computer science research. We outlined specific technologies for sharing data, software and knowledge, and emphasized the importance of legal and team/project internal agreements (*e.g.*, about sharing hardware resources, image publications, authorship). Finally, the paper also conveyed the considerations for grantors and grantees about the legal aspects, sharing agreements, time commitments and personnel resources needed for successful collaborative efforts which should be reflected in efforts planned for the future.

About the authors

Michael Simeone earned his Ph.D. in English from the University of Illinois at Urbana–Champaign. He is currently the assistant editor for *American Literary History*, as well as a project manager for the Institute for Computing in the Humanities Arts and Social Sciences. His research interests include consumer electronics and global digital culture, postmodern fiction, visual media, research databases, and practices of mapping and geospatial representation.

Jennifer Guiliano recieved a Masters of Arts in History from Miami University (2002) and a Masters of Arts (2004) in American History from the University of Illinois before completing her Ph.D. in History at the University of Illinois (2010). She currently serves as the Associate Director of the Center for Digital Humanities, a Research Assistant Professor in the Department of History at the University of South Carolina, and a Center Affiliate of the National Center for Supercomputing Applications.
Direct comments to jenguiliano [at] gmail [dot] com

Rob Kooper received his B.S. in computer science from the University of Delft in the Netherlands in 1996 and his M.S. from Georgia Institute of Technology in 2001. He currently serves as a research programmer within the Image Spatial Analysis Group at the National Center for Supercomputing Applications. His research interests are in human–computer interaction and graphics.

Peter Bajcsy has earned his Ph.D. degree from the Electrical and Computer Engineering Department, University of Illinois at Urbana–Champaign, Ill. in 1997, and M.S. degree from the Electrical Engineering Department, University of Pennsylvania, Philadelphia, Pa. in 1994. He is currently with the National Center for Supercomputing Applications at the University of Illinois at Urbana–Champaign, working as a research scientist on problems related to automatic transfer of image content to knowledge. Dr. Bajcsy's scientific interests include image processing, preservation, novel sensor technology, and computer and machine vision.

Acknowledgements

The project gratefully acknowledges the support of the National Science Foundation (NSF) and National Endowment for the Humanities (NEH) from the United States, and the Joint Information Systems Committee (JISC) from the United Kingdom in the form of a Digging into Data Challenge Grant Award.

The authors would like to thank Michael Meredith, University of Sheffield, Anne D. Hedeman, University of Illinois, Justine Richardson, Michigan State University, and Marsha MacDowell, Michigan State University of the DID–ARQ Project for their editorial contributions to this article. We also extend our gratitude to the complete DID–ARQ team: Peter Ainsworth, University of Sheffield; Steve Cohen, Michigan State University; Wayne Dyksen, Michigan State University; Kevin Franklin, University of Illinois; Karen Fresco, University of Illinois; Matt Geimer, Michigan State University; Anil K. Jain, Michigan State University; Robert Markley, University of Illinois; Amy Milne, Alliance of American Quilts; Dean Rehberger, Michigan State University; and, Tenzing Shaw, University of Illinois.

Notes

1. For example, the popular "Log Cabin" pattern has numerous common names depending on the arrangement, such as Streak of Lightning, Barn Raising, or Sunshine and Shadows.

2. For a complete list of sponsors of the Online Froissart, please visit http://www.hrionline.ac.uk/onlinefroissart/.

3. For a complete list of sponsors of the Quilt Index, please visit http://www.quiltindex.org/about.php#funding.

4. Medici, 2010; see Figure 2.

5. Apache™, 2010; labeled as svn in Figure 2.

References

Apache™ Subversion®, at http://subversion.apache.org, accessed 12 January 2011.

Gregory Crane, 2006. "What do you do with a million books?" *D–Lib Magazine*, volume 12, number 3, at http://www.dlib.org/dlib/march06/crane/03crane.html, accessed 16 December 2010.

Cyberintegrator, at http://isda.ncsa.uiuc.edu/cyberintegrator/, accessed 12 January 2011.

DID–ARQ Web site, at http://projects.matrix.msu.edu/did/, accessed 4 October 2010.

Digging into Data Challenge Competition Web site, at http://www.diggingintodata.org, accessed 24 January 2011.

Flickr, at http://www.flickr.com/, accessed 12 November 2010.

iRODS, at https://www.irods.org/index.php, accessed 14 November 2010.

Kepler, at https://kepler-project.org/, accessed 10 October 2010.

Medici, at http://medici.ncsa.illinois.edu/, accessed 12 November 2010.

Panoramio, at http://www.panoramio.com/, accessed 10 October 2010.

Taverna, at http://www.taverna.org.uk/, accessed 11 October 2010.

TinEye, at http://www.tineye.com/, accessed 14 December 2010.

TinyPic®, at http://tinypic.com/, accessed 10 December 2010.

Versus, at http://isda.ncsa.illinois.edu/versus/, accessed 12 January 2011.

Editorial history

Received 25 January 2011; accepted 15 April 2011.

Copyright © 2011, *First Monday*.
Copyright © 2011, Michael Simeone, Jennifer Guiliano, Rob Kooper, and Peter Bajcsy.

Digging into data using new collaborative infrastructures supporting humanities–based computer science research
by Michael Simeone, Jennifer Guiliano, Rob Kooper, and Peter Bajcsy.
First Monday, Volume 16, Number 5 - 2 May 2011
http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/printerFriendly/3372/2950

September 20, 2004

Don't Let Meetings Rule!

It's Monday morning and, if you are like most managers, you will spend a significant portion of your week in meetings. Yuck and double yuck for us all! I can remember when I landed my first salary job. The Farah Fawcett look was in and I thought going to meetings was cool. That was two decades ago, and I now loathe them - or most meetings anyway. Every manager I know feels the same. Most business meetings are a waste of time.

If no one wants to go to meetings and they aren't effective, why do we continue to book and attend these workplace gatherings of mental torture? Who knows....

Some meetings solve problems or generate ideas, but the other 95% GOTTA GO!

1. Just say no! Don't go to meetings that are not a good use of your time. I do this much more than I used to and LOVE IT! Every time I am NOT in a useless meeting, I feel a sense of pleasure and freedom. My goal is to improve on this and cut out even more.

2. If you must attend a meeting (and we all must), make it as quick, painless, and effective as possible. Here are a few tips:

Schedule meetings at least a week in advance

Here's a secret: When you plan and announce meetings a week or more in advance, you come across as organized and on top of things. This may not be true, of course, but for some reason having meetings scheduled out will leave this impression on others. So for that reason alone, it is a good habit! Scheduling a meeting in advance has a couple other advantages. First, you are more likely to find a time that works for everyone. Second, people will begin thinking about the subject (ok, maybe one person will think about it). To encourage participant thinking, include an explanation of the meeting topics and goals in your meeting notice.

Create and send out meeting agendas

This is another habit that will make you look good. Meetings need to have an agenda. When they don't they can become verbal free-for-alls where the most relentless and annoying people win (you've been there, right?). The agenda defines the topics that you want to cover, and the overall goal for the discussion. Agendas help you control the focus of your meeting.

Show up on time to meetings

Whether they say it or not, it annoys people when you are late to meetings, especially if you are always late. Is your time more precious than others? If you are often late to

meetings, chances are you are late to other appointments as well. Get over it and show up on time!

Send prereading materials to meeting participants

This has two purposes. First, to put together the prereading materials, you have to plan the meeting and know what you will cover. Planning the prereading helps you prepare for the meeting. Second, the prereading assignment will improve the quality of your meeting discussions.

Assume no one did the prereading and take the first few minutes to summarize

You ask: Why bother with the prereading if no one is going to take the time to review what I send out? Again, prepping the prereading is a good organization habit for you and the participants that don't ignore your assignment. Even if everyone reviews the prereading, it is still a good idea to summarize the material and why you thought it was important.

Bring food!

Sounds silly, but your attendance (physical and mental) will go way up if you serve doughnuts. Make it Krispy Kremes and everyone will show up for the meeting early or on time for fear of missing out! To get the biggest bang for your doughnut dollar, you will want to let people know there will be melt in your mouth bundles of sugar joy the day before (and again right before the meeting). Have fun!

Ask the right questions

Have you ever been in a meeting and had something to say but didn't because you didn't want the meeting to go any longer? Count on that this will occur. Closed-ended questions, or no questions, can make a meeting dull and unproductive. By asking provocative and relevant open-ended questions, you can wake up participants who may be zoning out and improve the likelihood that something good and interesting happens. If you are not going to ask for feedback or input, send an email or memo – don't waste people's time unless you want them to be a part of the conversation.

Encourage candid participation

When people wimp out at meetings, they become tiresome and useless – a waste of time! The purpose of a meeting is to have an open and productive conversation about an important business topic and to move the topic forward. Reaching this goal requires open and candid communication. Use facilitative phrases like, "what I hear you saying is.." or "I get the sense that you think this is not working.." or "I would like to know all the reasons you think this plan could fail."

Have fewer meetings!

I know this is redundant to #1, but it is worth mentioning again. Don't be a bad meeting maker. There are too many meetings. Push back on meetings that do not make sense or are unproductive. Only book a meeting when you have an important business issue to discuss and you want or need input, approval, or agreement. Even then, resist the urge to invite everyone and their brother – don't waste people's time unnecessarily. I am not a fan of regular staff meetings either - most are a waste of human resources and a mental drain.

Keep the meeting on track and on time

If you call the meeting, you call the shots. Respect people's time by keeping the meeting as short as possible and within the time period you communicated. Exceed everyone's expectations by finishing the meeting early! It is important to allow for participation but beware of this turning into scope creep. For example, a meeting focused on creating an IT wish list should not turn into a discussion of Bill Gates' gargantuan electricity bill.

http://managementcraft.typepad.com/management_craft/2004/09/dont_let_meetin.html

Feb 13, 2006

Avoiding Vampire Meetings

As a disclosure, I hate meetings. They are soul sucking, productivity draining wastes of time. However, they are necessary and a predicate of macro-productivity. Meaning: they halt your productivity, but are essential for a companies. Therefore, they are necessary for you, and day after day, week after week, you suffer through them only to emerge on the other side with a headache and a pile of voicemail.

Meetings at my office are no different. We try, mind you, to have a good meeting environment. Tasteful mid-century modern sofas, a plasma screen, a Starbucks machine. Hell, even cupcakes from Sprinkles! But still, I go into a meeting armed with ammunition to fight the need to scream (PPC-6700 and maybe a sketchbook) and leave feeling like an empty shell of a person. And I decided to fight back.

Why are meetings bad:

1) Its Not Me Its You Factor (INMIYF): Meetings usually are a way to offset blame. Meetings typically involve people who don't necessarily work together, but rely on each other, or report to a common boss. Hence, they provide an easy point of transmission of blame, as after a meeting is over, you return to separate corners.
2) The Kaplan Family Dinner Factor (KFDF): like my family, everyone thinks that what they have to say is really important. So they say it, regardless of who is also speaking, and at what volume they need to speak to get heard.
3) Digital Identity Maintenance Factor (DIMF): Exemplified by everyone on their Crackberry's, relentlessly maintaining their digital identities while their real world one languishes under the strain of market statistics.
There are more factors, but those are the three that bug me the most. Meetings start fine, as status updates or just getting things out there. Then they go into each phase in turn until it disintegrates into lunch. That is two hours gone, never to return.

How to Fix Then

So last week I tried to mitigate these factors. I only succeeded on two of them, but it worked. Since lists are the things that get the diggs, here you go:
1) Mitigate Blame Through Constructive Comments (MBTCC): When people start the finger pointing, take the opposite approach. Don't blame people, just use constructive questioning and socratic techniques to guide the conversation back to "how can this be fixed" from the "what the fuck went wrong" territory. You have to do this slowly, especially with different types of people. Some people are comfortable in blame mode, and taking them out of it can cause explosions. With those types, offeset the blame on a non-existent entity (I choose to blame marble, computers or made up words).
2) Talk Quietly, Quickly And Carry Big Words (TQQACBW): I call this the Professor Farnsworth factor. If you talk quietly, slowly and use words like "mitigate," "extrapolation," and other words not commonly heard in normal vernacular, people are forced to listen to you, as they think you are saying something Really Important. I do this often. I also speak pretty fast and my hair is all spiky and I wear black rim glasses, so I like to think I look nebbish and important. Or just geeky. Whatever it is, speaking quietly works. I've heard this technique works on little kids who get all yelly too.
3) Bring food! Cupcakes work.
4) Lastly, Come With an Agenda (CWA!): This is really important. I found that the best way to mitigate the Crackberry factor was to fill everyone's hands with a self-important piece of paper. It didn't mean anything, mind you. In fact the title was "Tech Meeting Agenda Type Thingy," and it had just bullets of things I talked about. However, people

like pretty letters and sans-serif fonts, so a piece of paper with both on it transfixed them enough so that the Crackberry's were put away, and their vacant looks went toward a piece of paper that might, given the right circumstances, mean something.

So there you have it! My recipe for meetings. Tomorrow at 11:30AM, we have another one, and I shall use these techniques to make sure that the meeting is fruitful, informative, and positive in outcome. Think of this as the silver bullet, the holy water and cross for the vampires.

[tags]Productivity, meetings, work, how-to[/tags]

http://www.blackrimglasses.com/2006/02/13/avoiding-vampire-meetings/