# Wrangling Your Data

# Preferred File Formats

Non-proprietary
Open, documented standard
Common usage by research community
Standard representation (ASCII, Unicode)
Unencrypted
Uncompressed

# Preferred Types

PDF/A, not Word
ASCII, not Excel
MPEG-4, not Quicktime
TIFF or JPEG, not GIF or JPG
XML or RDF

# Data Identifiers

PURL -- A PURL is a Persistent Uniform Resource Locator. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client.

DOI -- A DOI (Digital Object Identifier) is a name (not a location) for an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks.

# Data Identifiers

URI -- A Uniform Resource Identifier (URI) consists of a string of characters used to identify or name a resource on the Internet. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols.

# Back Ups

Make 3 copies
(e.g. original + external/local + external/remote)

Have them geographically distributed
(local vs. remote depends on recovery time needed)

## Common Back Up Types:

External Hard Drive
Tape Back Up
Cloud Storage: Amazon S3 or Glacier

# Back Up Testing

Test quarterly random files
Test full install annually
All three types of back up
Verify using forensics for bit-flip etc.

# Data Repositories

Github
Sourceforge
TAPAS (TEI Data)
All three types of back up
Verify using forensics for bit-flip etc.

# Migration/Transferability

Update regularly

Document any changes

Keep a copy of major versions of your site:
Internet Archive

Never build on your own server,
unless you plan to keep the site forever

# Hosting

Never build on your own server,
unless you plan to keep the site forever

Faculty move regularly: 3-5 years

Resources change regularly

# Websites

HTML v CMS

Who needs access?
Do you want to make updates for people?

Consistent elements/branding/design