

SAA Program, August 6, 2004
Digital Preservation Policies: Technical Considerations
Andrea Goethals

HANDOUT

A. What are SOME of the technical considerations that can be addressed in a preservation policy?

1. Applicable whether you are preserving your own files or someone else is preserving your files:

- what file formats will be used as access/preservation formats (sub-format restrictions?)
 - access formats = formats that you can use temporarily to display the content
 - preservation formats = formats you plan to support for a very long time (can use to generate new copies in access formats)
 - provide guidance/requirements (range of rigidity)
 - sub-format characteristics = data compression, color space, data encryption, character encoding, natural language specification, etc.
 - what preservation activities will you obtain/provide?
 - preserve the bits as-is ('bit-level' preservation) using media refreshment
 - migration to new successor formats
 - conversion to 'standard' formats
 - emulation
 - which files will receive which preservation activities?
 - based on file format?
 - based on selected content?
 - how will digital content be prepared for preservation?
 - unit(s) of content (groups of files, single files, etc.)
 - accompanying metadata (metadata format and required contents)
 - how to transfer/acquire content (FTP, DVD, etc.)
 - how to package (directory structure, naming scheme, file compression allowed?)
 - virus-checked, etc.
 - digitally signed?
 - encrypted?
 - recorded message digest by data owner! (so that you know that your original files haven't changed)
-

2. Applicable if you are preserving your own or someone else's files:

- how will files be stored? (media, store repeats)
 - Database, tape, hard drive, etc.
 - Global files used? (commonly-used files like XML schema stored once and other packages can point to these files)
- what metadata will be collected/stored and which parts by humans/computer programs?
 - per format, per unit of content

- authority of internal vs external metadata (do you trust the metadata contained within files over the metadata submitted along with files, at least for certain file characteristics?)
- metadata format for metadata files created after the files are in the archive (METS, etc.)
- how metadata will be stored and updated (database, tape, hard drive, tied to content (stored same place as archived files), combo)
 - redundant metadata (ex: in Database and in flat files stored with the submitted files?) - trade-offs: more to maintain but more robust
- will the original submitted files be kept as-is forever?
- how digital content will be identified (naming scheme) in the archive
 - implications for the scalability of your archive - maximum number of 'archivable' files per day or per some other time unit based on your naming scheme
- how digital content (and its media) will be monitored for integrity? (how digital content owners will know that their digital content hasn't changed)
 - redundant files (number of stored copies, different geographical locations)
 - frequency of media refreshment
 - frequency and extent of integrity checks
 - specific and redundant message digest algorithms
- how 'distributed' content will be handled (links, schema)
 - localize or rely on external stability of linked-to files?
 - localize = make a new copy of the file and replace all external links with the identifiers of locally-stored files
 - which links to follow (some links are more important than others)
 - ex: important links: XML schema, XML style sheets, CSS
 - how far to follow the links?
 - ex: do we care about the files linked-to from the files that this file links to?
- how digital content can be retrieved/accessed/withdrawn
 - unit(s) of content
 - versions
 - access requirements

B. URLS mentioned in my presentation:

FCLA Digital Archive (Policy Guide, Recommended formats, Action Plans):

<http://www.fcla.edu/digitalArchive/daInfo.htm>