

WebWise 2007 – Pre-Conference Session #1 – Preserving Digital Collections

Preserving Digital Collections – Priscilla Caplan (FCLA) – Florida Center for Library Automation - pcaplan@fcla.edu/digitalArchive

- 1) Curation, Archiving and Preservation
 - a. forces thinking about lifecycle management – proactive management (Creation, Appraisal, Documentation, Reuse)
 - b. Creation: choice of format (<http://www.digitalpreservation.gov/formats> - LC Sustainability of Digital File Formats)
 - c. Creation: use of format (<http://www.niso.org/framework/Framework2.html> - A Framework of Guidance for Building Good Digital Collections)
 - d. Documentation: descriptive metadata, source and intellectual property rights, project-wide details (who, what, when, how), persistent identifier, digital provenance
 - e. Selection/Appraisal/Review – digital is different (impact of abundance, more weight to practical considerations, earlier decision points; can't wait or may lose digital file), **focus on high risk/high consequence** (<http://www.dca.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/appraisal-and-selection.pdf> - Digital Curation Manual, Appraisal and Selection)
 - f. Use and Reuse – dynamic data continues to change, implication is now thinking of systems rather than individual digital objects – feeds into research and publication process; may have secondary publications that result from original and those need to be preserved
- 2) Digital Preservation Basics
 - a. The process of ensuring that a digital object is usable over the long term
 - b. Archiving and Preservation
 - i. Availability – you can't preserve what you don't physically control. Methods: web harvesting, institutional repositories, contracts with suppliers, deposit agreements
 - ii. Identify/understandability – ties to descriptive metadata
 - iii. Authenticity – source and content must be verifiable; maintaining complete event history and chain of custody, documentable fixity
 - iv. Fixity – quality of not being unintentionally altered or deleted: methods for ensuring fixity include: sound storage management, media refreshment, and calculating and checking checksums
 - v. Viability – quality of being readable from media; threatened by media degradation and media obsolescence. Methods to ensure viability: media refreshment, media migration.
 - vi. Renderability – quality of being displayable or otherwise usable. Threatened by digital file format obsolescence. Methods: reformatting, maintaining old hardware and software, programs written emulating old hardware and software.
 - c. Rights – Preservation rights scenarios from Karen Coyle, “Rights in the PREMIS Data Model” (<http://www.loc.gov/standards/premis/Rights-in-the-PREMIS-Data-Model.pdf>)

- d. Economics – Not just question of money, wider issue of incentives – parties who have need (beneficiary), right (rightsholder), or ability (the archive) to preserve; issues that impede preservation (beneficiaries are not rights holder, difficulty of separating high and low-end services) - Brian Lavoie, “The Incentives to Preserve Digital Materials” – <http://www.oclc.org/research/projects/digipres/incentives.dp.pdf>
- 3) Preservation and Practice
- a. Conceptual Framework (OAIS) – what does it mean to be OAIS compliant?
 - i. Reference model – <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1/pdf>
 - ii. Framework for understanding and applying concepts needed for long-term preservation of digital information; gives common vocabulary; models for what an OAIS must do
 - b. Trusted Digital Repositories
 - i. Trusted Digital Repositories: Attributes and Responsibilities (2002:RLG and OCLC) <http://www.rig.org/legacy/longterm/repositories.pdf>
 - ii. May be locally run or a third party, but must accept responsibility for the long-term maintenance of digital resources on behalf of depositors and for the benefit of current and future users
 - c. Progress toward repository certification
 - i. *RLG/NARA Repository Certification Audit Checklist* (first draft)
 - ii. CRL Certification of Digital Archives project
 - iii. Repository certification metrics – ISO standardization
 - d. Preservation Strategies (from Thibodeau 2002) – emulation, version migration, etc. (many different options)
 - e. Preservation Metadata – PREMIS Data Dictionary – defines core preservation metadata pertaining to objects, agents, events, and rights
 - f. Significant Properties – what properties should be retained if item is recreated in the future through migration or emulation? (i.e. intellectual content only, links, formatting, look and feel, behaviors)
 - g. Preservation infrastructure (set of tools to help with preservation, all are moving targets) – standards and best practices, file format registries (GDFR), environment registries, technology watch services, large scale storage management, distributed network of trusted repositories
 - h. Repositories and preservation systems
 - i. Third party (OCLC digital archive)
 - ii. Open source (DSpace, Fedora, LOCKSS, DAITSS – Florida uses this; developed with IMLS funding)
 - iii. Vendor (DigiTool)
 - iv. Custom (in-house, contract)
 - v.

Preserving Government and Political Information: the Web-at-Risk Project – Valerie Glenn

- Web Harvesting – what and why?
 - o Automated capture of web materials in danger of disappearing

- Capture a particular event or moment in time
- Build a collection of similar or related materials
- Web-at-Risk grant funded by NDIIPP
 - purpose: to build tools to allow librarians to “capture, curate, and preserve web-based government and political information”
 - sample collections: CyberCemetery (U. of North Texas); UCLA Online Campaign Literature Archive; Islamic and Middle Eastern Political Web (Stanford)
- Web Harvesting – issues involved
 - building collections and tools to capture those collections; identification of content; depth of capture; number and frequency of captures; permissions
- Harvesting Tools: based on Heritrix (<http://crawler.archive.org>); HTTrack (<http://www.httrack.com/>); Web Curator Tool (<http://webcurator.sourceforge.net/>)
- Harvesting Services: ArchiveIt! (<http://www.archive-it.org>); OCLC Digital Archive (<http://www.oclc.org/digitalarchive/default.htm>); Web Archiving Service (development in progress through Web-at-Risk project)
- Web-at-Risk wiki: <http://wiki.cdlib.org/WebAtRisk/tiki-index.php>
- NDIIPP: <http://www.digitalpreservation.gov/>

Public Television: Preserving Digital Programs – Mary Ide

- NDIIPP PTV Project – goal is to design a digital environment that will store files in packages and put them in an open-source storage application that will be maintained and refreshed
 - Delivery to LC
 - Partnership with NYU (repository design for ingest and delivery)
- PBS distribution (common metadata standard; PREMIS standard; standard definition digital beta, DSpace to manage files with Fedora on front end, OAIS compliant)
- Digital preservation – ideal format is not here yet for video
- Book for Reference: *Digital Video Preservation Reformatting Project* – Mellon Foundation supported project

The Romance of Lost Causes: Preserving Digital Art – Rick Rinehart, UC Berkeley Art Museum/Pacific Film Archive

- very complex needs for archiving “born digital”
- “Archiving the Avant-Garde” project (NEA grant – research and development project ended last December) – findings:
 - At first project seemed technical – save computers? No, but need to save physical/esthetic object – is the art in the computer? No, but yes it is, too.
 - What is it we want to preserve?
 - Do we want to preserve or keep it alive?
 - Can’t just preserve code, it is not robust enough for art objects; so developed a new language/documentation for composing scores for digital art (media art notation system – XML – code complex digital art objects) – Los Alamos
 - Digital art will be an interpretation of the art; no longer a “master” work, more socially-created digital art; emphasizes behavior over form
 - Causes museums to rethink preservation (butterfly display to butterfly hut for breeding and observing)

