Challenges of Digital Preservation:

Early Lessons from the Portico Archive

Eileen Fenton
Executive Director, Portico

iPRES 2007
Beijing, China
10-11 October 2007

Issues and Questions

1. What is digital preservation?

2. Case study: An overview of Portico

3. Early lessons from operational experience

Digital Preservation Means …?

- Reformatting from print to digital to create access surrogate or expand product line
- Byte storage only without regard to ongoing usability
- Assuring enduring content usability and access
  - "The managed activities necessary for ensuring both the long-term maintenance of a bit stream and continued accessibility of content" **From *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report, May 2002.*
  - Ensuring the "usability of a digital resource, retaining all qualities of authenticity, accuracy, and functionality deemed to be essential"
  - *** From *The Preservation Management of Digital Material Handbook* maintained by the Digital Preservation Coalition in collaboration with the National Library of Australia and PADI

Components Necessary for Preservation

- “Urgent Action”* statement suggests preservation is distinct from access and requires a “minimal set of well-defined services”
- Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC) produced by the Center for Research Libraries identifies requirements
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) toolkit released by the Digital Curation Centre and DigitalPreservationEurope provides self assessment tool

* “Urgent Action Required to Preserve Scholarly E-Journals”
http://www.arl.org/bm~doc/ejournalpreservation_final.pdf

Components Necessary for Preservation

- An organizational mission that highlights the preservation function
- A sustainable economic model able to support preservation activities over the targeted timeframe
- A robust and regularly updated technological infrastructure able to support an identified preservation strategy and best practices
- Clear legal rights
- Relationships with content providers/creators and (eventual) users

Portico's Mission

To preserve scholarly literature published in electronic form

and to ensure that these materials remain available

to future generations of scholars, researchers, and students.

Portico's History

- In 2002, JSTOR initiated a project known as the Electronic-Archiving Initiative, the precursor to Portico.
- The goal was to facilitate the community's transition to secure reliance upon electronic scholarly journals by developing a technological infrastructure and sustainable archive able to preserve scholarly e-journals.
- Portico was launched in 2005 by JSTOR with support from Ithaka, The Andrew W. Mellon Foundation and the Library of Congress.
- Portico is a not-for-profit organization with a mission and singular focus to provide a permanent archive of electronic scholarly resources, beginning with e-journals.

Portico's Approach: Content Scope

In scope:

- Initially electronic scholarly, peer reviewed journals Priority is given to publishers or titles recommended by librarians
- Intellectual content of the journal, including text, tables, images, supplemental files
- Limited functionality such as internal linking

Out of scope:

- Full features and functionality of publisher's delivery platform
- Ephemeral look and layout of today's HTML rendition of a journal

Portico's Approach: Migration Supplemented with Byte Preservation

- Publishers deliver "source files" of electronic journals (SGML, XML, PDF, etc.) to Portico.
- Portico converts proprietary source files from multiple publishers to an archival format suitable for long-term preservation. Portico's preservation format is based on the NLM Archiving DTD.
- Source and normalized files are deposited in the archive. Once deposited, content must remain in the archive.
- 2 million+ articles or 34 million+ files are archived to date; 1.7 million articles are available for audit/verification viewing
- Portico migrates files to new formats as technology changes.

Portico's Approach: Access

- Portico offers access to archived content to only those libraries supporting the archive.
- Access is offered only when specific trigger event conditions prevail **and** when titles are no longer available from the publisher or other sources.
- Trigger events initiate campus-wide access for all libraries supporting the archive regardless of whether a library previously subscribed to the effected content.
- Libraries may rely upon the Portico archive for post-cancellation access, **if** a publisher chooses to name Portico as one of the mechanisms designated to meet this obligation.

Sources of Support

- Support for the archive comes from the primary beneficiaries of the archive.
- Contributing publishers supply content and make an annual financial contribution (USD $250 to $75,000).
- More than 6,200 journals from more than 40 publishers are committed to the archive to date.
- Libraries make an Annual Archive Support (AAS) payment based upon total library materials expenditures (USD $1,500 to $24,000).
- More than 375 libraries from 9 countries are “Archive Founders.”

Emerging Lessons: Publishers

- Publishers understand the library market now demands robust preservation arrangements.
- Publishers want to be a part of the archiving solutions that libraries support.
- Publishers are developing multi-layered strategies.

Emerging Themes Lessons: Libraries

- Libraries are actively evaluating the scope of their archival responsibilities and options for meeting these.
- Multi-layered strategies responding to library needs to preserve a wide array of e-content are beginning to emerge.
- Breadth of archival strategy varies with institutional size.
- Coordinated e-preservation strategies and print collection management strategies are being developed.

Emerging Lessons: Archive Operations

- Like digital preservation, electronic publishing is still evolving. Best practices are still emerging.
  - Publishers' platforms, formats and data structures are shifting.
  - E-journal content is now frequently in 3 formats: current e-formats, early e-formats and digitized print.
  - Archives must be prepared to respond to – and influence - the complexity of this still shifting landscape.

Emerging Lessons: Archive Operations

- Journal content is complex – and not always tidy.
  - Online journals may be the product of multiple data streams. Complexity increases opportunities for errors in content.
  - Publisher systems are oriented toward on-time publication. Preservation issues are not the focus.
    - PDF validity may vary
    - Early e-issues reveal various production issues.
  - Publishers are open to input about how to create publications that are more easily archived. Archives can play a consultative role.

Emerging Lessons: Archive Operations

- Gathering and communicating holdings information is challenging.
  - Publishers do not have readily available publication histories or inventories.
  - Gathering and reporting detailed, accurate holdings data in a machine-readable way is difficult.
  - Absent solid holdings data, affirming completeness of the archive is difficult.
  - This is a fruitful and important research area.

Emerging Lessons: Archive Operations

- File usability vs. validity creates special challenges.
  - Files may be usable to a reader but not technically valid. Creates special format migration concerns.
  - Files may be technically valid but not usable to readers.
  - Helpful to tackle this issue while content creator can participate in the resolution.
  - This issue impacts digital repositories of all types.
  - A good area for collaborative tool development.

Emerging Themes and Lessons: Implications

- Libraries should continue to make their preservation needs and preferred strategies known to publishers.
- Publisher and library preservation strategies benefit from being informed by one another.
- Overt communication of archival strategies or intentions assists both parties.

Eileen Fenton
eileen.fenton@portico.org
www.portico.org