

Welcome
to the
Portico Participants' Meeting
ALA, New Orleans
June 24, 2006

8:00 – 8:30 a.m. Breakfast

8:30 – 10:00 a.m. Program



Portico

Publisher Relations Update

Toni Tracy

Director, Portico Publisher Relations

Portico Participants Meeting

ALA, New Orleans

June 24, 2006



Publisher Relations Update

- Thirteen publishers have signed agreements to participate in Portico
- Discussions underway with an additional 64 publishers across scholarly publishing – commercial publishers, not-for-profit societies, university presses
- Existence of Portico and other archiving entities has resulted in a new conversation in the scholarly publishing community around the question: What is our archival strategy?



Portico Participating Publishers (as of 6/20/05)

- American Anthropological Association
- American Mathematical Society
- Annual Reviews
- Berkeley Electronic Press
- BioOne
- Elsevier
- John Wiley & Sons
- Oxford University Press
- SAGE Publications, Inc.
- SIAM
- Symposium Journals
- UKSG
- University of Chicago Press



Title Update (as of 6/20/05)

- Number of Journals Committed to Portico = 3,558
- Archive operations are “live” and work has begun to ingest content from signed publishers
- Number of Articles from AMS, Berkeley Electronic Press, OUP, and Wiley ingested into the Portico archive = more than 20,000



Toni Tracy
toni.tracy@portico.org
www.portico.org



Portico

Library Relations Update

Ken DiFiore

Associate Director, Library Relations

Portico Participants Meeting

ALA, New Orleans

June 24, 2006



Library Relations Update

- Greater awareness of e-journal preservation issues.
- Increased dialog about archive strategies.
- Community response to Portico has been outstanding!
- 100 committed libraries.
- > 100 more libraries expressed interest.
- Starting outreach to consortia and international communities.



Ken DiFiore
ken.difiore@portico.org
www.portico.org



Issues in Archiving Electronic Journals

Evan Owens
Chief Technology Officer

Portico Participants Meeting
ALA, New Orleans
June 24, 2006



Preservation of Digital Objects

- Ensuring long-term viability
- 20, 50, 100 years from now, can we
 - read the files?
 - understand the structure of the files?
 - be sure that we have an authentic copy of the work?
- Layers
 - Physical Layer: storage media
 - Logical Layer: file formats, structured data
 - Conceptual/Intellectual Layer: the “work”
- Approaches to preservation:
 - Emulate (or maintain) the original technology
 - Migrate (and/or normalize) to currently supported formats
 - Byte preserve for future digital archeologists



Digital Preservation Prerequisites

- Content
- Metadata
 - Descriptive (e.g., author/title; “who”)
 - Technical (e.g., file formats; “what”)
 - Administrative (e.g., rights; events)
- Standards, file formats
 - Legal, open, de facto, proprietary, ...
- Standards watch:
 - Key activity of an archive
 - Migration before obsolescence
 - Requires expertise in relevant standards and technologies
 - Likely genre-specific



Varieties of Digital Preservation Projects

- Library and media digitization projects
 - Controlled environments; potential for good metadata
- Web site harvesting
 - Uncontrolled environment; minimal metadata available
- Electronic records retention
 - Potential for lots of control; mandatory metadata and formats
- Published electronic content
 - Semi-controlled; good descriptive metadata; variable or no technical metadata
- Scientific data
 - Enormous quantities of data
 - High expectations for long-term usability



Electronic Journals and Digital Preservation

- Journal publishing models are evolving
 - Publishing practice varies:
 - Print only, E-only, both
 - More / less / same in each edition
 - E-product varies:
 - HTML Header & PDF
 - HTML Full-text with links and supplemental stuff & PDF
 - HTML only
- A “work” with multiple “manifestations”
 - XML or SGML source files
 - Print PDF used to drive printing press
 - Web PDF optimized for online delivery
 - HTML header or full text (often generated from XML or SGML source)



Portico Archival Strategy for E-Journals

- Source file archiving
 - Preserve the components not the rendition
 - Include high-resolution files (PDF and figures) if available
 - All e-only components (data, media, etc.)
 - SGML / XML structured text by preference
 - HTML as last resort
- Preserve intellectual content not “look and feel” of HTML
 - HTML renditions are an artifact of current technology
 - Often dynamically generated
 - Fragile technology, overdue for change
- Preserve only essential features of the user interface
 - Reference linking, other content-based features
 - Not generic navigation or search or e-commerce features



Portico Preservation Implementation

- Key technical influences:
 - GDFR, PREMIS, METS, MPEG-21, ARK, OAIS
- Format-based migration strategy
- Preservation policies:
 - Fully supported
 - Reasonable effort
 - Byte-preserve only
- Preservation policies based on
 - Format validity
 - File format action plans and archive capabilities
 - Business rules such as publisher preference
- Archive must preserve supporting information
 - Required files such as DTDs and entity files; Documentation; Contracts



Portico Technical Infrastructure

- Content processing and archive systems
 - Documentum, Oracle, Sun Solaris, Sun & Hitachi storage
 - Currently housed at Princeton University OIT
- Delivery system
 - Managed by JSTOR, currently located at Princeton University
- Offline data replication 2006-2007
 - Multiple copies to “hard media” for distributed storage.
 - Media will be a mix of DVD and hard disk.
 - Locations in North America and one in Europe.
 - Storage providers will be both commercial and academic.
- Online data replication 2007-2008
 - Online replication with synchronized mirror sites
 - In addition to offline replication



Electronic Journal Data Issues

- Inputs
 - Per article: one text or metadata file, zero or more other files
 - Arbitrary (publisher-specific) collections of data
 - Proprietary file & directory naming conventions
 - Standard and/or Proprietary formats for text and metadata
 - Undocumented business rules hidden in the data
- Outputs
 - Content normalized to NLM Archive and Interchange DTD
 - Metadata: technical, descriptive, events
 - Packaged in Portico METS
- Portico DTD 2.0 extends NLM DTD 2.1
 - All added text tracked with markup:
`<x x-type="archive">(added text)</x>`



Data Normalization Strategy

- “Archive” not “aggregate” or “re-publish”
- Don't lose data
- Don't add data tacitly
 - Additions are marked using <x> tag
- Preserve the publication, not the business process
 - E.g., discard initials of copy editor or proof mail date
- Preserve semantics of publisher markup
 - Even if apparently incorrect
- Don't second guess the publisher
- Resolve all publisher-specific rules during normalization
 - E.g., mapping of external file names to XML structures
- Recognize that publisher practices change over time



Problem Areas in Current E-Journal Publishing Practice

Based on our evaluation of publisher data

- Content management and quality control
 - Documentation, naming, packaging
 - Production content: PDF, XML, graphics
 - Author-supplied supplemental content: various formats
- Structured metadata and use of persistent identifiers
 - Must be able to cite and link to online edition
 - DOI or equivalent persistent link
- Versions and revisions
 - Differences between renditions (HTML, PDF, print, XML/SGML)
 - Policy regarding and tracking of revisions and updates
- Issue-level content
 - Covers, front matter, back matter



Evan Owens
evan.owens@portico.org
www.portico.org



Watch for...

- Announcements of new participating publishers
- News of additional library Archive Founders
- Account information for participating libraries
- Additions to the Portico website



Thank you for your attention.

Always feel free to contact us.

participation@portico.org
www.portico.org

