


I T H A K A



Content Production for Current and Future Reuse:
Insights from the Archiving and
Digital Preservation Communities

Evan Owens
Chief Technology Officer
Ithaka Electronic-Archiving Initiative

STM E-Production Seminar
2nd December 2004

I T H A K A

Abstract

In this era of multi-channel, multi-product publishing, production processes must produce data for multiple uses including interchange and reuse. Archiving or “digital preservation” can be seen as another kind of reuse, one with a longer time span. Current activity in the library and archival communities can shed some light on issues in the organization of STM content that will make content work better for immediate needs and for the longer term.



I T H A K A

Five Visions of Electronic Publishing Circa 1996

What does electronic publishing mean?

Author — “No more proofreading. Immediate publication of my work!”

Publisher — “Production tasks can be automated.”

Librarian — “Cheaper serials, more complex operations.”

Reader — “Access to the full literature for free!”

Archivist — “Everything will be lost!”

— Boyce, Biemesderfer, and Owens, *Serials Review*, 1996

We hope that not all of these visions of the future are correct!



What Archiving Is

- Long-term preservation
- 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 archiving:
 - Emulate (or maintain) the original technology
 - Migrate (and/or normalize) data to supported formats
 - Byte preservation (for digital archeologists)



What Archiving Is Not

- “Open Archive Initiative”
- Making backup copies or mirror sites
 - Necessary but not sufficient
 - What the IT industry calls “archiving”
- Writing files to tape and storing it in a salt mine
 - Necessary but not sufficient
- Publishing (and vice-versa)
 - But there is overlap
- Aggregation (and vice-versa)
 - But there is overlap



What Archiving Requires

- Content
- Metadata
 - Descriptive
 - Intellectual identify of the work
 - e.g. author, title, journal, volume, page
 - Structural
 - How the files go together to make a representation of the intellectual object
 - Technical
 - Information about the files
 - Administrative
 - Rights information
 - History of the objects in the archive



Archiving Projects and Metadata

Some representative projects:

- Library digitization of physical objects
 - Photos, manuscripts, rare books, etc.
 - Controlled environment; good metadata
- Web site harvesting
 - Uncontrolled environment; minimal metadata
- Electronic records retention
 - Government and business
 - Potential for required metadata
- Published electronic content
 - Good descriptive metadata; variable or no technical metadata



Ithaka Electronic-Archiving Initiative

- Trusted third-party archive for libraries and publishers
- Source file archiving of Electronic Journals
 - Not web renditions per se
 - Including print or high-res source files whenever possible
- SGML / XML when available; headers if not
 - Normalize to standard XML DTD for long-term maintenance, not as value-add
 - HTML as last resort
- Get content into archive as cost-effectively as possible
 - Minimal intervention
 - “Archive” not “aggregate” or “re-publish”
- Identify formats, validate, and characterize
 - For future migrations as necessary
- Still in pilot stage; launch in early 2005
 - Comments based on sample data from 10 pilot publishers



PreMIS www.oclc.org/projects/pmwg/

- Preservation Metadata Implementation Strategies
 - OCLC/RLG-sponsored international working group representing libraries, archive, museums, and other interested parties
- Preservation metadata
 - Information necessary for the processes that support the long-term retention and accessibility of digital materials
- PreMIS objectives:
 - Implementation survey
 - “Core” metadata element set and data dictionary
- Why is it important?
 - Digital information is notoriously fragile
 - At risk from bit decay and technological obsolescence
 - Preservation activities can be needed soon after creation
 - Much sooner than for most physical objects



E-Journals as Archival Objects

- E-Journal articles can have multiple manifestations
 - Print rendition
 - Online HTML rendition
 - Online PDF rendition
- And a variety of source materials
 - Print PDF or Postscript
 - Print graphics (high-res)
 - Web PDF, HTML, inline images, graphics
 - Full text or header files (SGML or XML)
 - Often never delivered directly to reader
 - Media files (e.g., video, sound)
 - Other electronic stuff (data, software, etc.)
 - Often called “supplemental”
- How do all the pieces go together?
- What is the “work”? What needs to be preserved?



I T H A K A

Structural Metadata: What is to be Archived?

- No industry standard for e-journal content packaging or transmission
- Some real-world problems:
 - Wide variety of naming conventions
 - Between publishers, journals, even issues
 - Absence of manifest and sequence information
 - Mismatch between headers and PDFs
- Archive must know what each file is and how they go together as machine-readable data
 - And so should publishers!
 - Not as implicit business rules or practices (next slide)
- Adopt a naming convention, document it, and then apply it rigorously
- Industry standard required ASAP!



Structural Metadata: Hidden Business Rules

- Problems in mapping SGML/XML to files
- Examples
 - `<display-formula id="df1">` means look for a graphic called `df1.gif`
 - `<figure filename="fig1">` means look for a set of files called `fig1_t.gif`, `fig1_m.gif`, `fig1_h.tiff`
- Obscure or convoluted mapping rules are fragile!
- A well-defined “interface agreement” is needed
 - Between the data and the data delivery system
 - Between the data today and systems years from now
- Even when the data is moving internally, pretend that you are dealing with an external partner
 - Assume reuse; the rest follows
- Good interfaces make good neighbors



I T H A K A

Data Exchange: How to get it from here to there

- Library of Congress (NDIIPP) has funded projects experimenting with moving data from one archive to another
 - And back again...that's the real test!
- Publishers are increasingly facing the same problem
 - Content is reused by third parties
 - Exporting metadata (e.g., abstracts)
 - Exporting entire content sets including full text
 - To other systems
 - As journals change publishers
 - As delivery systems are upgraded and replaced
- Key is rigorous control over the content
 - Clear interface agreements
 - And some industry standards



Data Exchange: Versions / Revisions

- Version control problems
 - SGML / HTML / PDF mismatch
 - Web site corrected in place without updating source files
 - Pieces missing from SGML, only in HTML
- Revision policies vary
 - Replacement
 - Annotations in place
 - Watermarks
 - Appended errata
- A data exchange or syndication problem:
 - How to broadcast adds, changes, and deletes to the archive
- A business practices problem:
 - Publishers must know what you have, what you have changed, and where the changes are
 - Have a policy about revisions to online publications and enforce it



Technical Metadata: Formats, Dependencies

- What formats ought to be used?
- Digital preservation community argues over merits of particular file formats
 - Are they “open” enough or “archival” enough?
 - Begins to be theological
- Most e-journal data is fine
 - Graphics formats are usually not a problem
 - But make sure that all fonts are embedded
 - Avoid Microsoft proprietary formats
 - Consider adoption of PDF/A when it becomes available
- Badly coded SGML/XML is as evil as Quark or MS-Word
 - My opinion



Technical Metadata: Formats, Dependencies

- What format is this file?
 - And any embedded formats
- Projects in archival community:
 - Global Digital Format Registry (GDFR)
 - Digital Library Federation sponsored project
 - PRONOM (UK)
- Format information key to preservation by migration
- Everyone needs to identify formats
 - Especially DTD versions for XML / SGML
 - Declare the DTD and DTD version in the document instance
- Preserve documentation
 - Especially DTD documentation
- Identify “dependencies”
 - Such as DTD and entity files
- All good business practices for publishers too!



I T H A K A

Technical Metadata: Significant Characteristics

- Printing/Publishing industries are experts on many formats
- Library/Archiving community sometimes reinvents the wheel
- Preservation community experimenting with graphics migration such as from JPEG to JPEG 2000
- Discovering that file format alone is not sufficient; they need what are called “significant characteristics”
 - E.g., in migrating a collection of photos of coins, optimize the conversion to emphasize contrast rather than color
- It depends on human judgment of intellectual content
- Sounds a lot like publishing to me!



Other Technical Metadata Elements

- Environment Information
 - Hardware / software required to render or execute a file
- Fixity / Authenticity / Integrity
 - Checksums, digital signatures
- Creating Application
- Inhibitors
 - E.g., copy protection, passwords, DRM
- Encoding
 - E.g., compression, tar, binhex, base64
- Derivation
 - Source from which file was created
- Publishing works from the default environment of today
- Archives have to assume that the environment is going to change significantly



Events Metadata

- The event history of an object
- Audit trail, as it were
- CMS provide this automatically
- Journal Publishers tend to have fixed workflow patterns
 - Received, edited, proofed, published
 - Less demand for audit trails
- Archives are more like records retention systems
 - Must demonstrate exactly what was done when and by whom so as to prove authenticity and provenance
- Doesn't hurt publisher to think about this, but probably not worth serious investment



Rights Metadata

- An area where archives follow the lead of publishers
- Easy case: Publisher has all rights to the content
 - And can authorize the actions necessary for archiving
- Difficult case: Publisher has restricted rights to some content
 - Must be able to express this in a machine-actionable format
 - A non-trivial task
- Rights expression languages are challenging
- Capturing rights information is even more challenging
- Archives and publishers have work to do in this area



Descriptive Metadata

- The easy case: header file and PDF
 - No question what is metadata and what is the “published” information
 - Many archival projects fit this model; e.g., photo collections
- The harder case: SGML/XML full text
 - The text is also source of the metadata
 - Functional schizophrenia
 - Is this the front matter as printed on the page?
 - Or is it the database entry for this unit of content?
- An extreme example:
 - A publisher who repeats the entire journal title change history in every journal article
- Massive data redundancy



Descriptive Metadata, continued

- In an archive, descriptive metadata has to be held external to the preserved object
 - So that it can survive a migration of the object
 - So that it can be edited or annotated without changing the preserved object
- E-Archive solution
 - Extracted metadata (from header or full text)
 - Curatorial metadata (copied, then edited as needed)
- Publisher solutions?
 - Please don't lose track of the text as published
- Publishers who run secondary services have lots of experience with this problem



Summary

- An archive must have metadata that includes
 - Bibliographic description (who)
 - Manifest of pieces (what)
 - Rights information (if not blanket) in machine readable form
 - Control of differences between versions and over time
 - Technical information about file formats
 - Especially any unusual formats
- So must publishers...sooner or later
- Bottom line recommendations
 - Good basic content management practices, well-executed
 - Common sense on policy issues such as supplemental materials
 - Recognize that something things will only be byte-preserved
- Easier said that done, of course
- Devil is in the details . . .



I T H A K A



Content Production for Current and Future Reuse:
Insights from the Archiving and
Digital Preservation Communities

Evan Owens
Chief Technology Officer
Ithaka Electronic-Archiving Initiative

STM E-Production Seminar
2nd December 2004