Starting a Digital Preservation Program

Illinois Library Association
October 12, 2007

Sarah Shreeves
Tim Donohue
Tom Teper
Outline

• What is Digital Preservation Management?
• What is IDEALS?
• Meeting Our Preservation Commitment
• Implementing DPM Activities within IDEALS
• Next Steps
What is Digital Preservation?

- Digitization
- Using “archival” CDs
- Collecting electronic records
- Building an institutional repository
- Running back-ups

Digital Preservation is better understood as a management process.
Definitions

• Preservation - Providing access to materials for as long as they are needed by whomever needs them.
  • Includes: Policies and Procedures, Preventative Preservation, Collections Conservation, Conservation, Reformatting, Replacement…

• Digital Preservation Management - Process that requires the use of the best available technology as well as carefully thought out administrative policies and procedures to maintain access to materials.
  • Includes: Organizational Concerns, Technology Implementation and Management, Resource Management
Digital Preservation Management

- The process of building a platform that will enable an institution to maintain access to its collections.
  - Can be:
    - Locally developed solutions
    - Locally implemented commercial/open-access solutions
    - Contractually secured
    - Outsourced
  - Typically is:
    - All of the above
The Foundation I

- The Open Access Information System (OAIS) Reference Model
The Foundation II

- Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC) v. 1.0
  - Originally Developed by the RLG/NARA Digital Repository Certification Task Force
  - Evaluated by the Center for Research Libraries
  - Expanded and Revised from the Original Draft

- Does not require OAIS compliance, but draws very heavily from the OAIS Reference Model
The Framework I

- Organizational Framework -
  - The policies, procedures, practices, people—the elements that any programmatic area needs to thrive, but specialized to address digital preservation requirements. It addresses this key development question:
    - *What* are the requirements and parameters for the organization's digital preservation program?
The Framework II

- Technological Infrastructure -
  - Consists of the requisite equipment, software, hardware, a secure environment, and skills to establish and maintain the digital preservation program. It anticipates and responds wisely to changing technology. It addresses this key development question:
  - How will the organization meet defined digital preservation requirements?
The Framework III

- Resource Framework -
  - Addresses the requisite startup, ongoing, and contingency funding to enable and sustain the digital preservation program. It addresses this key development question:
    - *What* resources will it take to develop and maintain the organization’s digital preservation program?
The Developmental Stages

- **Acknowledge** – understanding that digital preservation management is a local concern
- **Act** – Initiating digital preservation management projects
- **Consolidate** – Segueing from projects to programs
- **Institutionalize** – Incorporating the larger environment and rationalizing programs
- **Externalize** – Embracing inter-institutional cooperation
<table>
<thead>
<tr>
<th>Stage</th>
<th>Rank</th>
<th>Organizational</th>
<th>Technology</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledge</td>
<td>1</td>
<td>Non-existent, implicit, very high level</td>
<td>Non-existent, heterogeneous,</td>
<td>Generally low, finite, ad hoc financial commitments</td>
</tr>
<tr>
<td>Digital preservation is a local concern</td>
<td></td>
<td></td>
<td>decentralized</td>
<td></td>
</tr>
<tr>
<td>Act</td>
<td>2</td>
<td>Implicit or general, increased evidence of commitment</td>
<td>Project-specific, reactive, ad hoc</td>
<td>Often project-based funding</td>
</tr>
<tr>
<td>Initiate digital preservation projects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consolidate</td>
<td>3</td>
<td>Basic and essential policies</td>
<td>Assess technology investment, more proactive</td>
<td>Some funding and support beyond projects, but limited</td>
</tr>
<tr>
<td>Segue from projects to programs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institutionalize</td>
<td>4</td>
<td>Consistent, systematic, comprehensive policy framework for planning</td>
<td>Anticipate needs, investments defined by management, implemented across system</td>
<td>Sustainable funding identified for core program areas and enhancement</td>
</tr>
<tr>
<td>Incorporate the larger environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Externalize</td>
<td>5</td>
<td>Virtual organizations complement institutions; collaboration inherent in resource planning</td>
<td>Distributed and highly integrated; extra-organizational features/services</td>
<td>Varying levels of investment, but sustained funding; possibly distributed management</td>
</tr>
<tr>
<td>Embrace collaboration and dependencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Digital Preservation Platform
Getting Started with Digital Preservation

- Start with a discrete, manageable collection of content
- Start with materials that you have a mandate to preserve - whether by tradition or by project scope
- Start with the understanding that it will be an ongoing, evolving process.
What is IDEALS?

Institutional Repository for the scholarship and research in digital form of the faculty, students, and staff as well as material that reflects the intellectual environment of the University of Illinois at Urbana-Champaign.

Joint project of CITES and the University Library and supported by the Office of the Provost.

http://ideals.uiuc.edu/
IDEALS

- Systematic dissemination of deposited works
- Preservation
- Persistent and reliable access

- Focused on working with a handful of early adopters

- Pilot phase currently winding to an end
  - Shift to production likely at the start of 2008
What type of materials?

Also audio and video
What is an institutional repository?

A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members.

It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.

Clifford Lynch, Executive Director
Coalition for Networked Information
In the beginning: Promises, promises

- Can we really commit to preserving everything?
- What does it really mean to preserve this stuff?
- What kind of staff expertise do we need?
- What kind of resources do we need?
- What kind of technical infrastructure do we need?
Getting our act together

• Got our Preservation Librarian involved

• Training and self education
  • Cornell’s Digital Preservation Management Workshop and Online Tutorial
    http://www.library.cornell.edu/iris/tutorial/dpm/eng_index.html
  • Understanding Open Archival Information System conceptual model
  • Trustworthy Repositories Audit Checklist
Takeaways:

- You do need to be explicit about what you will do and what you won’t do.
- You don’t have to preserve everything if you say you aren’t.
- Digital preservation management is not about the technology.
Establish pilot policy

http://www.ideals.uiuc.edu/about/IDEALSPreservationSupport.html

- Repository Support

Policy is realistic and feasible for where we were.
Getting our act together, cont.

Backup tapes stored next to the server!

Not Really Our Server Room!

Photo by Sylvar. Used under a Creative Commons 2.0 Attribution license. http://www.flickr.com/photos/sylvar/
Looking forward to production: Digital Preservation White Paper

http://hdl.handle.net/2142/135

- Laid out for the Library and CITES administration what supporting a digital preservation management program would mean:
  - **Commitment on the part of both organizations**
  - **Resources** in terms of funding and staff are specifically allocated
  - Processes, policies, and the institutional commitment are documented and as transparent as possible.
  - The technical infrastructure is developed using community standards.
  - Commitment of resources for planning and community standards building.
IDEALS Preservation Policy: Organizational Framework and Commitment

https://services.ideals.uiuc.edu/wiki/bin/view/IDEALS/IDEALSDigitalPreservationPolicy

- **Mandate**
  - Agreement that we are making with our user community
  - Role of the University Library in preserving access to material

- **Objectives**
  - Persistent access
  - Trusted service for our user community

- **Scope**
  - Research and scholarship

- **Who’s responsible?**
  - CITES and the Library
IDEALS Preservation Policy: Operating Principles

- Compliance with the Open Archival Information System (OAIS) Reference Model standard.
- Adherence to prevailing community standards for preserving access to digital content whenever possible.
- Participation in the development and implementation of standards.
- Commitment to an interoperable, scalable digital archive with appropriate storage management for content.
- Policies, procedures, and practices are clearly documented and consistent.
- Maintains hardware, software, and storage media containing archival content in keeping with prevailing best practices.
- Establishes procedures to meet archival requirements pertaining to provenance, chain of custody, authenticity, and integrity.
- Complies with intellectual property, copyright, and ownership rights for all content.

- Aiming for compliance with certification requirements for a Trustworthy Repository.
What resources do we need?

- Funding
  - Currently from the Office of the Provost
- Designated staff
  - Built into our job descriptions

Technology infrastructure

- Move from Library to CITES
  - Better environment
  - Better security
  - Distributes support for the tech infrastructure
Risks and Challenges

- Technological Change
- Sustainability
- Partnership between the University Library and CITES
- Identifying an Exit Strategy
Moving towards actionable policies and procedures
Putting the Plan into Practice

- Policy should lead Technology (not vice-versa)
- “Support” Policies *will* change
  - Reassessment necessary
  - Document decisions…
  - and reasons!
- “Best Practices” – no reason to go it alone
What will IDEALS “support”?  

- What have others done?  
  - Michigan’s Deep Blue – Preservation & Format Policy  
  - Florida Digital Archive – Policies & Format “Action Plans”  
  - Library of Congress – Sustainability of Formats  
  - Australian Partnership for Sustainable Repositories (APSR)  

- What Support Policies are we missing?  
  - Digital Preservation Support Policy  
  - Format Support “Matrix”  
  - Format Recommendations
Digital Preservation Support

- Format-based Categories of Support
  - **High Confidence**
    - Full Support (including migration)
  - **Medium Confidence**
    - No migration promised
  - **Low Confidence**
    - “Bit-level” support only

![Diagram showing categories of support]

- Openly Documented
- Widely Adopted
- Widely Supported
- Uncompressed or Lossless Compression
- No Embedded Content or DRM

*(size ≠ weight)*
Format Support Matrix

- Compilation of “known” formats
- Concentration on textual formats

<table>
<thead>
<tr>
<th>Proprietary</th>
<th>Microsoft Office</th>
<th>OpenOffice.org, HTML</th>
<th>Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited Adoption</td>
<td>OpenOffice.org</td>
<td>Microsoft Office, HTML</td>
<td>Widely Adopted</td>
</tr>
<tr>
<td>Limited Support</td>
<td>Microsoft Office</td>
<td>Adobe PDF, HTML</td>
<td>Widely Supported</td>
</tr>
<tr>
<td>Embedded Content / DRM</td>
<td>MS Powerpoint (w/ Audio or Video)</td>
<td>MS Powerpoint</td>
<td>Nothing Embedded</td>
</tr>
<tr>
<td>Lossy Compression</td>
<td>JPEG</td>
<td>TIFF, JPEG 2000</td>
<td>No/Lossless Compression</td>
</tr>
</tbody>
</table>
Format Recommendations

Textual

↑ CSV, Text, PDF/A, XML*
Open Document Format
↓ RTF, MS Office, PDF, HTML

Audio

↑ AIFF, WAVE, Ogg Vorbis, FLAC
↓ AAC, MP3, Real, WMA

Images

↑ TIFF, JPEG 2000
↓ GIF, JPEG, PNG

Video

↑ AVI, Motion JPEG 2000
↓ MP2, MP4, Quicktime, WMV

High Confidence / Preference
Medium Confidence / Preference
What we are doing

- Basic Activities (All Items: 🔝 ↔️ ↓️)
  - Regular Virus Scans, Checksum verification
  - Nightly off-campus backups
  - Refresh storage media
  - Preservation Metadata (minimal)
    - Format, checksum, file size, etc.
  - Permanent Identifiers (Handles)
  - *Always* keep the original document
  - Monitoring and reassessment of formats
    - Very minimal/infrequent for ↓️
What we are doing

- Intermediate Activities (↔)
  - Additional monitoring, more frequent reassessment
  - When possible, attempt to migrate formats to preserve *content* and *style* (hopefully)
    - No promises that *functionality* will be preserved
    - (e.g.) Powerpoint → PDF (*possible functionality loss*)
    - (e.g.) PDF 1.4 → PDF/A (*possible style loss*)
What we are doing

- Full Support Activities (⬆)
  - Additional monitoring, more frequent reassessment
  - When necessary, migrate document to successive format.
  - Attempt to preserve *content, style* and *functionality*
    - (e.g.) PDF/A → successor to PDF/A
Policy → Technology

OpenOffice.org Format Converter

Before

After

(e.g.) http://hdl.handle.net/2142/2364
Our First Problem…

- Character issues in Word (and PDF)
- Found by chance
- Consultation with submitter
- Originally Wordperfect
- Re-submitted as RTF

A White and Nerdy:
Computers, Race, and the Nerd Stereotype
Lori Kendall

Previous research on nerds has analyzed the relationship of this stereotypical identity to issues of race, gender, and computer expertise. For instance, in an earlier article, I argue that narratives such as that presented in the popular movie, Revenge of the Nerds, depict the incorporation of the previously marginalized nerd identity into closer alliance with hegemonic masculinity, demonstrating the increasing legitimacy of expertise in computers as a form of masculine prowess. However, I also suggest that the continued negativity of the nerd stereotype reveals a persistent uneasiness with computer use and computer users (285). In a similar analysis, Ron Eglash analyzes images of nerds as white and male by default, yet hardly a portrait of white male superiority (80). He explores possibilities for reversal, analyzing images of black nerds in popular culture, and attempts by black and women to subvert the nerd stereotype. However, ultimately he notes that nerd is still used in the pejorative sense; its routes to science and technology access are still guarded by the unmarked signifiers of whiteness and male gender (60).

Both of these articles point to contradictions in nerd identity that allow it to both maintain normative boundaries of power and offer sites for intervention (Eglash 49). It is logical to expect that the tension inherent in these contradictions would resolve over time, as computers become more ubiquitous in society. We’ve been through several cycles of developments in computer- and Internet-related technologies, including the phenomenal wave of internet start-ups in the 1990s (just prior to my previous article) and the subsequent dot-com bust of 2000 (just prior to Eglash’s). In the U.S., information technologies are increasingly part of most people’s lives. As the household presence of computers becomes no more extraordinary than that of other consumer electronics such as televisions and microwave ovens, one might expect that the nerd stereotype would fade from view, an anachronism from an earlier age, reflecting now-defunct uneasiness with the then-new computer technologies.
What we are NOT doing

• Checking every file for content problems
  • (e.g.) character encodings, DRM, embedded content

• Verifying ALL automated migrations are “successful”

• Checking validity of format (e.g. JHOVE)

• Removing/modifying/replacing original file
  • Exceptions: viruses found or OCR necessary
Next Steps - Policies

- UIUC Library following our lead
  - Digital Preservation Management program
  - Library “Best Practices”
- IDEALS Preservation Working Group
  - Internal analysis using Trustworthy Repositories Audit & Certification (TRAC) checklist
    - Policy/Procedure Gap Analysis
  - Ongoing format reassessment
  - “Human-understandable” Policies/Procedures
Next Steps - Implementation

- Additional automated migrations
- Additional preservation metadata
- Check validity of formats (e.g. JHOVE)
- On upload virus scanning
- Best practices / technology monitoring:

GDFR - Global Digital Format Registry

The technical registry: PRONOM

DROID - Digital Record Object Identification

Xena - Digital Preservation Software

JHOVE
Contact Information and References

- Tom Teper – tteper@uiuc.edu
- Sarah Shreeves – sshreeve@uiuc.edu
- Tim Donohue – tdonohue@uiuc.edu
- IDEALS: http://www.ideals.uiuc.edu/
- IDEALS Wiki
  - http://services.ideals.uiuc.edu/wiki/
  - Information Section → “Policies”
  - Working Groups Section → “Preservation”