Day-to-Day Digital Preservation: A Case Study

GSLIS Data Curation Institute
June 4, 2008

Tim Donohue
IDEALS Technical Lead
Outline

• Intro to IDEALS
• How we got started...
• Our preservation strategies
• Current / future data curation work
• Next steps...
What is IDEALS?

Digital repository for the scholarship and research of the faculty, students, and staff of the University of Illinois at Urbana-Champaign.

- Dissemination
- Persistent Access
- Preservation

http://ideals.uiuc.edu/

A joint initiative between the University Library and CITES with support from the Office of the Provost.
What type of materials?

Also audio and video
IDEALS goals:

- Help increase access to published and unpublished research
- Help increase the impact to published and unpublished research
- Provide a persistent, permanent URL for citing research
- Preserve research for long term access and use
The IDEALS “service model”

- Not just a repository...
- *Set of services* to collect, disseminate, and provides persistent and reliable access to the research and scholarship of faculty, staff, and students at the University of Illinois at Urbana-Champaign.
Other services we offer…

- Consultation on copyright issues
- Access restrictions / Embargo of items
- Statistics on number of downloads (working on departmental monthly reports…)
- Pilot service to deposit research into PubMed Central and ‘disciplinary’ repositories
IDEALS: the beginning
In the beginning: Promises proceed us

- 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.icpsr.umich.edu/dpm/
  • Understanding Open Archival Information System (OAIS) conceptual model
  • Trustworthy Repositories Audit & Certification (TRAC)
The Digital Preservation Platform

Borrowed from the ICPSR Digital Preservation Tutorial:
http://www.icpsr.umich.edu/dpm/
Preservation Takeaways:

- 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 just about the technology.

Photo borrowed from: http://flickr.com/photos/santos/
Getting our act together, cont.

Not Really Our Server Room!

Backup tapes stored next to the server!

Photo borrowed from: 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: Operating Principles

Adherence To:
- OAIS Reference Model
- Community standards for preserving digital content
- Hardware, software, and storage media best practices.
- Intellectual property, copyright, and ownership rights of all content

Commitment To:
- Interoperable, scalable digital archive
- Clear, openly documented policies & procedures
- Archival requirements for provenance, custody, authenticity, integrity

Goal: A Certified, 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
How IDEALS supports data (files)
What have others done?

- Michigan’s Deep Blue – format support policy
  - [http://deepblue.lib.umich.edu/about/](http://deepblue.lib.umich.edu/about/)
- Florida Digital Archive – format “action plans”
  - [http://www.fcla.edu/digitalArchive/](http://www.fcla.edu/digitalArchive/)
- LC: Sustainability of Digital Formats
  - [http://www.digitalpreservation.gov/formats/](http://www.digitalpreservation.gov/formats/)
- Australian Partnership for Sustainable Repositories (APSR)
  - [http://www.apsr.edu.au/](http://www.apsr.edu.au/)
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

---

**Openly Documented**

**Widely Adopted**

**Widely Supported**

**Uncompressed or Lossless Compression**

**No Embedded Content or DRM**

Low Confidence (gray area)

(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

Data Concentration

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 (extremely minimal)
    - Format, checksum, file size, etc.
  - Permanent Identifiers (Handles)
  - *Always* keep the original file(s)
  - Monitoring and reassessment of formats
    - Very minimal/infrequent for
What we are doing

• Intermediate Activities (↔)
  - Automated nightly “access copies” generated for major formats
  - When possible, attempt to migrate formats to preserve content and style (hopefully)
  • No promises that functionality will be preserved

Examples:
MS Excel → CSV (possible functionality loss)

MS Word → PDF (possible style / font loss)
What we are doing

• Full Support Activities (↑)
  - When necessary, migrate document to successive format.
  - Attempt to preserve *content, style* and *functionality*

Example:
OpenOffice.org 2.x → OpenOffice.org 3.x
What we are NOT doing

- Checking every file for content problems
  - 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
Making data available in IDEALS
Data in IDEALS: still early days

- Mostly ‘simplistic’ data sets

- Data in spreadsheets, text, XML formats
  - i.e. “familiar” formats

- Problem: capturing relationships (between datasets, procedures, papers, etc.) in DSpace
Vole Demographic Data

Lowell Getz (Dept of Animal Biology)

- 25 years of data
- Data in 20 Excel files
- HTML Explanations
- PDF Manuscripts
- HTML “Sitemap” organizes data/files

http://hdl.handle.net/2142/161
Illinois long-term selection exp. for oil & protein in corn

Department of Crop Sciences

- Data since 1896 (ongoing)
- SAS Statistical System files (text)
- ReadMe describes experiments
- Tech Reports in PDF
- Collection description organizes data/files

http://hdl.handle.net/2142/3524
Soon: Crystallography Data

Scott Wilson, School of Chemical Sciences

- Processed data in CIF (Crystallographic Info File)
- Transform to CML (Chemical Markup Lang.)
- Original unprocessed data kept on server in Clark X-Ray Facility (size concerns)
- Data available to SPECTRa Search tools…

Borrowed from Jmol website:
http://jmol.sourceforge.net/

http://www.lib.cam.ac.uk/spectra/
Future: Morrow Plots Data

- Oldest continuous agricultural research fields in USA (est. 1876), 2nd in world
- National Historical Landmark
- Data ongoing
- Likely require digitization of lab notebooks, etc.

Gaps – What we are NOT doing

- Making data “useable” directly from IDEALS
- Making data itself “searchable” (besides metadata)
- Providing unique “visualizations” of data
  - Some will be coming, for crystallography data

Photo borrowed from: http://www.flickr.com/photos/cseesze
Sustainability issues…

- How do we preserve *more* than just files?
- How do we keep data understandable?
- Disk space concerns… versus access

Photo borrowed from: http://www.flickr.com/photos/columna
A more “ideal” future

- When possible, finer grained access to data
- Better ways to show relationships between data and results/papers
- Begin to develop models for talking to faculty about their data

Photo borrowed from: http://www.flickr.com/photos/dsevilla/
For More Information

http://ideals.uiuc.edu/

Sarah Shreeves  
IDEALS Coordinator  
sshreeve@uiuc.edu

Tim Donohue  
Technical Lead  
tdonohue@uiuc.edu

Policies / Documentation:

http://services.ideals.uiuc.edu/wiki/