3-3-2010

Issues in Digital Preservation

Denise Troll Covey
Carnegie Mellon University, troll@andrew.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/lib_science
Part of the Other Social and Behavioral Sciences Commons

Published In
School of Library and Information Science, University of Pittsburgh.

This Article is brought to you for free and open access by the University Libraries at Research Showcase @ CMU. It has been accepted for inclusion in University Libraries by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Issues in Digital Preservation

Denise Troll Covey
Principal Librarian for Special Projects
Carnegie Mellon

University of Pittsburgh
School of Library and Information Science
March 3, 2010
Presuppositions

• Digital preservation presupposes a digital copy

• A digital copy presupposes knowing copyright law and the copyright status of the work
  – Determining © status is no easy feat
  – See Peter Hirtle, “Copyright and the Public Domain in the United States”

http://copyright.cornell.edu/resources/publicdomain.cfm
Making a digital copy

• If out-of-copyright, can digitize without permission

• If in-copyright, can digitize
  – If © owner grants permission
    • Presupposes knowing identity and location of © owner and getting a response from © owner
  – Or if © law grants exemption allowing digital copy to be made without permission of © owner
  – Otherwise digitizing is © infringement
    • Sanctions apply, including statutory damages
Exemptions (for eligible libraries / archives)

- **Preservation copies**
  - 108(b) allows 3 copies of unpublished work in the library / archive collection
  - Copies can be deposited in other libraries / archives
  - Digital preservation copies cannot be used

- **Replacement copies**
  - 108(c) allows 3 copies of published work that is damaged, deteriorating, lost, stolen or in obsolete format
  - *If an unused copy is not available at a fair price*
  - Digital replacement copies must be used on-site
Section 108 Study Group

- March 2008 recommended allowing
  - Preservation copies of published work
  - Preservation copies of public online content
    - Immediate access on-site; embargo on remote access
  - Making replacement copies from preservation copies
  - Replacement copies of fragile work
  - Remote access to digital replacement copies
  - Outside contractors to digitize © material
    - If contracted by eligible library or archive, cannot retain copies or benefit commercially from the copies

- Recommendations not yet sent to Congress
Acquiring © permission

- Permission to digitize and provide open access

<table>
<thead>
<tr>
<th>Random sample feasibility study</th>
<th>Rare books study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Books</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Total items</td>
<td>337</td>
</tr>
<tr>
<td>1. Copyright protected</td>
<td>94%</td>
</tr>
<tr>
<td>2. © owners contacted</td>
<td>88%</td>
</tr>
<tr>
<td>3. © owners responded</td>
<td>51%</td>
</tr>
<tr>
<td>4. Permission granted</td>
<td>43%</td>
</tr>
<tr>
<td>Overall success rate</td>
<td>22%</td>
</tr>
</tbody>
</table>

1999-2001                      | 2003    |
Analysis by rights holder type

Random sample feasibility study

Rare books study
Rare book study permission costs

$78 per book/volume

$ 10,808  FTE labor
$   379  Phone calls
$   100  Paper & postage

$ 11,287  TOTAL

May 2003 – October 2003

Does not include legal fees, administrator time, or cost of Internet connectivity or database creation.
Metadata

• Needed for long-term semantic interoperability
  – Descriptive metadata
    • Bibliographic and copyright information

• Needed for long-term preservation
  – Technical metadata
    • How content is encoded and structured, e.g., software format and version
  – Administrative metadata
    • Aspects of the capture process, e.g., scanning date, scanner profile, lighting
Preserving digitized work

• Follow standards and best practices for scanning
  – Master files in open standard format, e.g., TIFF
  – Descriptive, technical and administrative metadata

• Use loss-less compression to store master files

• Keep redundant copies in different locations

• Check and maintain integrity of files

• Refresh files and media (tapes, discs) periodically

• Options as formats develop and change
  – Migration – can render / deliver work – ideal
  – Emulation – can mimic original rendering
  – Preservation of bits – cannot render
Preserving born digital work

- Capture descriptive metadata
- Convert to standard delivery format if necessary
- Keep redundant copies in different locations
- Check and maintain integrity of files
- Refresh files and media (tapes, discs) periodically
- Problems as formats develop and change
  - No master files in open standard format
    - PDF is proprietary
    - No technical or administrative metadata
      - What version of PDF? Created how?
Institutional repositories (IR)

- Collect, preserve and provide open access to
  - Technical reports and working papers
  - Conference papers and presentations
  - Dissertations and theses
  - Teaching materials
  - Lectures (videos)
  - Journal articles
  - Books

- Exemptions in © law do not apply
  - Need permission of © owner
IR preservation policy

• Items will be retained indefinitely
  – Unusual file formats may not be readable or accessible over time

• Files will be backed up regularly
• Submission policy and procedures
  – Must deposit bibliographic metadata and full text
  – No embargo period allowed
  – Administrator vets submissions for
    • Eligibility of author / depositor
    • Compliance with scope and format of IR content
  – Authors / depositors are responsible for validity and authenticity of submissions and © violations
• Submission policy and procedures
  – Must deposit bibliographic metadata
  – Deposit of full text encouraged
    • Embargo period allowed
  – Administrator vets submissions for
    • Eligibility of author / depositor
    • Compliance with scope and format of IR content
    • Alignment with publisher policy
  – Authors / depositors are responsible for validity and authenticity of submissions
  – *Unclear who is responsible for © violations*
In theory

- Open access complies with © law
  - Green OA – retain the right to self-archive work on a website or repository
    - Slow voluntary participation ➔ OA mandates
  - Gold OA – publish in OA journal or hybrid journal and pay the requisite fee
    - Resistance: impact factor of OA journals, fees

In practice

- Much of what is available OA infringes ©
Green OA

- 63% of publishers allow open access
  - [http://www.sherpa.ac.uk/romeo/](http://www.sherpa.ac.uk/romeo/)

- Conditions and restrictions apply
  - Versions
    - Pre-print: Original Manuscript, Submitted Manuscript
    - Post-print: Accepted Manuscript, Version of Record
  - Embargoes
    - Typically 12 to 24 months after publication
    - Can be different for website and repository
    - Can be different for journal titles published by the same publisher
## Restrictions on version per venue

<table>
<thead>
<tr>
<th>Publisher policy</th>
<th>Original or Submitted Manuscript</th>
<th>Accepted Manuscript</th>
<th>Version of record</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pubs</td>
<td>Titles</td>
<td>Pubs</td>
</tr>
<tr>
<td>Allowed on website</td>
<td>43%</td>
<td>83%</td>
<td>65%</td>
</tr>
<tr>
<td>Allowed in IR</td>
<td>38%</td>
<td>75%</td>
<td>54%</td>
</tr>
<tr>
<td>Prohibited on website</td>
<td>48%</td>
<td>14%</td>
<td>33%</td>
</tr>
<tr>
<td>Prohibited in IR</td>
<td>52%</td>
<td>22%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Journals in which Carnegie Mellon faculty publish
Total publishers = 282  Total titles = 2,833
Examples Green OA maintenance

• Original or Submitted Manuscript
  – Allowed after acceptance for peer review
  – Allowed, but must be removed after acceptance for publication
  – Allowed, but must be removed after publication
  – Allowed on website, but must remove during peer review
  – Allowed on website after publication
  – Allowed on website, prohibited in IR
  – Allowed in IR, prohibited on website
  – Allowed in IR, but must be removed when submitted for publication

• Accepted Manuscript
  – Allowed after publication
  – Allowed after embargo
  – Allowed on website after publication; allowed in IR after embargo
  – Allowed on website after publication; prohibited in IR
  – Allowed on website after acceptance for publication; prohibited in IR
  – Allowed on website after embargo; prohibited in IR
  – Allowed in IR; prohibited on website
  – Allowed in IR after embargo; prohibited on website
Example policy details

- Publisher © and source must be acknowledged
- Must link to publisher version with statement that this is the definitive version and DOI
- Must state that version is the author’s version

© ACM, YYYY. This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in PUBLICATION, {VOL#, ISS#, (DATE)}
http://doi.acm.org/10.1145/nnnnnn.nnnnnn
Participation at Carnegie Mellon

77% of articles cited on faculty websites could be OA
32% of articles are OA

<table>
<thead>
<tr>
<th>College</th>
<th>Articles</th>
<th>Faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>OA</td>
</tr>
<tr>
<td>Business and economics</td>
<td>1,415</td>
<td>16%</td>
</tr>
<tr>
<td>Computer science</td>
<td>2,340</td>
<td>55%</td>
</tr>
<tr>
<td>Engineering</td>
<td>4,713</td>
<td>24%</td>
</tr>
<tr>
<td>Fine arts</td>
<td>200</td>
<td>17%</td>
</tr>
<tr>
<td>Humanities &amp; social studies</td>
<td>2,419</td>
<td>40%</td>
</tr>
<tr>
<td>Public policy &amp; management</td>
<td>380</td>
<td>26%</td>
</tr>
<tr>
<td>Science</td>
<td>3,414</td>
<td>41%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>14,881</td>
<td>32%</td>
</tr>
</tbody>
</table>

Study conducted 2007-08
## Alignment with publisher policy

38% of OA articles infringe publisher ©
18% of infringing articles policy prohibits OA

<table>
<thead>
<tr>
<th>College / Policy alignment</th>
<th>Not known</th>
<th>Aligned</th>
<th>Not aligned</th>
<th>Prohibited OA</th>
<th>Prohibited Pub PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business and economics</td>
<td>15%</td>
<td>67%</td>
<td>18%</td>
<td>7%</td>
<td>93%</td>
</tr>
<tr>
<td>Computer science</td>
<td>16%</td>
<td>53%</td>
<td>31%</td>
<td>5%</td>
<td>70%</td>
</tr>
<tr>
<td>Engineering</td>
<td>9%</td>
<td>56%</td>
<td>35%</td>
<td>25%</td>
<td>61%</td>
</tr>
<tr>
<td>Fine arts</td>
<td>18%</td>
<td>24%</td>
<td>58%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Humanities &amp; social studies</td>
<td>16%</td>
<td>21%</td>
<td>63%</td>
<td>9%</td>
<td>90%</td>
</tr>
<tr>
<td>Public policy &amp; management</td>
<td>26%</td>
<td>29%</td>
<td>45%</td>
<td>22%</td>
<td>76%</td>
</tr>
<tr>
<td>Science</td>
<td>6%</td>
<td>65%</td>
<td>29%</td>
<td>47%</td>
<td>53%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>12%</strong></td>
<td><strong>50%</strong></td>
<td><strong>38%</strong></td>
<td><strong>18%</strong></td>
<td><strong>73%</strong></td>
</tr>
</tbody>
</table>
Alignment with publisher policy

• Infringement would be much higher than 38% if assessed compliance with all policy details

• Could not assess compliance
  – Distinguish author manuscript versions
  – Determine if embargo was respected
  – Track maintenance

• Assessed alignment based on whether policy
  – Allowed open access
  – Prohibited, allowed or required the publisher PDF
My position: exercise and foster

- **Civil disobedience**
  - In a democracy, when conscience and law clash, we are morally justified, if not duty bound, to follow our conscience, not wait for the law to change.
  

- **Moral courage**
  - When opposed and success entails risk, we must act to preserve the values of honesty, fairness, respect, responsibility and compassion.

Outreach to faculty

- Benefits of OA: increased access and citation
  - Importance of retaining right to self-archive
  - Faculty Senate 2007 open access resolution
- Benefits of IR: OA; statistics; indexed by Google, professional maintenance; Selected Works gallery
- Library deposits work self-archived on websites or sent as attachments in email
- Library converts work to standard format
- Library provides how-to documentation

Silent phase: Nov 2008
Public push: June 2009
1,940 papers to date
22,794 downloads to date
## Articles

<table>
<thead>
<tr>
<th>Title</th>
<th>Journal/Source</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Archiving Journal Articles: A Case Study of Faculty Practice and Missed Opportunity</td>
<td>portal: Libraries and the Academy</td>
<td>2009</td>
</tr>
<tr>
<td>Using Data to Persuade: State Your Case and Prove It</td>
<td>Library Administration and Management Journal</td>
<td>2005</td>
</tr>
<tr>
<td>The Need to Improve Remote Access to Online Library Resources: Filling the Gap between Commercial Vendor and Academic User Practice</td>
<td>portal: Libraries and the Academy</td>
<td>2003</td>
</tr>
<tr>
<td>Copyright Permission: Turning to Dust or Digital</td>
<td>International Journal of the Book</td>
<td>2003</td>
</tr>
<tr>
<td>Academic Library Assessment: New Duties and Dilemmas</td>
<td>New Library World</td>
<td>2002</td>
</tr>
</tbody>
</table>

[View More >>](#)
Organizational support for IR

- Research Showcase Coordinator
  - Reports to Head of Archives and DL Initiatives
- Scholarly Communications Committee
  - Scholarly Communications Forum
- Libraries Council
- Faculty Senate Library Advisory Committee
- Liaison librarians
Resistance to IR

• Some liaison librarians
  – Complain that IR doesn’t support appropriate formats
  – Won’t invest time in learning how IR works
  – Not prepared to answer questions from faculty

• Some faculty
  – Afraid of not getting published if pre-print is in IR
    • Dean talked to the publisher: no basis for fear
  – Don’t like citation format of harvested work – so we fixed it
  – Don’t understand, need or want OA – need to educate
  – Don’t have time to OA – so library does it for them
Why so few contributions?

- Mismatch of vocabulary, incentives and workflows
  - E.g., faculty are loyal to discipline, not institution
- Lack of awareness
- Lack of reward
- Lack of time
- Attitude of peers
- Copyright

Rochester IMLS research (2005)
Rights and Rewards in Blended IRs (2006)
Many authors

• Don’t understand © or open access
  – Don’t want to lose control of their work
• Don’t know who owns © to their work
  – Don’t understand, read or keep © agreements
• Don’t have time or confidence to negotiate © transfer
• Are concerned won’t get published if self-archive
• Are concerned self-archiving breaks © agreements
• Ignore © agreements and self-archive

Incentives

- Likely to contribute to repository
- Much more likely to contribute

- Support
- Management / preservation
- Necessary for pay award
- Peers are contributing
- Mandate in institution
- Peer leaders are contributing
- Mandate in department
- Necessary for promotion
- Assert copyright

Rights and Rewards (UK, 2006)
Trusted digital repositories

• Mission: provide long-term access

• Approach: framework of attributes + responsibilities
  – Identify reliable, sustainable repository infrastructures
  – Form basis for development of trusted services

RLG-OCLC Report (May 2002)
RLG - OCLC recommendations

1. Develop framework and process to certify trusted digital repositories
   - Criteria that must be met
   - Mechanisms for periodic assessment

2. Research and create tools to identify the attributes of digital materials that must be preserved
   - Technical metadata needed
   - Tools to capture significant properties at creation
3. Research and develop models for cooperative networks and services
   – Intellectual versus technical aspects
   – Examples of service-level agreements

4. Design and develop unique, persistent identifiers
   – Design identifier systems for long-term maintenance, storage and access
RLG - OCLC recommendations

5. Investigate and disseminate information about relationship between digital preservation and ©
   – Understand how current © law inhibits preservation and how technical strategies impinge on ©
   – Develop tools to identify roles and responsibilities of
     • Content creators
     • Organizations that preserve content
   – Develop models to
     • Acquire copyright permission
     • Contract agreements (© owners and libraries / archives)
RLG - OCLC recommendations

6. Determine best technical strategies to provide continued access – understand
   – Technical + © implications of migration / emulation
   – How technical strategies affect need for metadata

7. Define minimal metadata required for long-term management of digital materials and develop tools to generate or extract as much as possible
   – Technical, administrative and descriptive metadata
Audit and certification criteria

- Organizational infrastructure
- Digital object management
- Technologies, technical infrastructure, security

CRL Report (February 2007)
http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

1. Develop framework and process to certify trusted digital repositories
   - Criteria that must be met
   - Mechanisms for periodic assessment

RLG-OCLC recommendations (May 2002)
Organizational infrastructure

- A1. Governance and organizational viability
- A2. Organizational structure and staffing
- A3. Procedural accountability and policy framework
- A4. Financial sustainability
- A5. Contracts, licenses, and liabilities
Digital object management

• B1: Initial phase of ingest (acquisition)
• B2: Final phase of ingest
• B3: Preservation strategies and mechanisms
• B4: Minimal conditions for performing preservation
• B5: Minimal-level metadata
• B6: Ability to produce and disseminate accurate, authentic versions of digital objects
Technologies, technical infrastructure and security

• C1: General system infrastructure requirements

• C2: Appropriate hardware and software and mechanisms to evaluate when changes are needed

• C3: Security
  – Technology (e.g., servers, firewalls, routers)
  – Fire protection
  – Flood detection
Example: Organizational infrastructure

• **A1. Governance and organizational viability**
  – Succession / contingency plans or escrow arrangements

• **A3. Procedural accountability and policy framework**
  – Publicly accessible definitions and policies dictating how preservation service requirements will be met
  – Documented history of changes in operations, procedures, software and hardware
  – Commitment to transparency and accountability

• **A4. Financial sustainability**
  – Short- and long-term business planning processes
  – Commitment to monitor and bridge gaps in funding
Google is not a trusted repository

- Proposed settlement is a catastrophic mistake that will create an environment that asphyxiates culture
  - Giving Google exclusive license to orphan books
  - Regulating and obsessively controlling access to books at the level of a page or a quote

- Cannot rely on favors granted by private companies to define access to our culture

Lawrence Lessig
Joi Ito CC license
Estimate 3.5 M orphan books, 13 M out-of-print books, and 5 M books presumed © owner does not respond

Based on random sample feasibility study
Thank you!

troll@andrew.cmu.edu