Repository models and policies for preservation

Steve Hitchcock

Preserv Project
Intelligence Agents Multimedia Group,
School of Electronics and Computer Science (ECS),
Southampton University

DPC Briefing on
Policies for Digital Repositories: models and approaches
Preserv preservation service provider schematic
OAIS functional model
Service provider model
Institutional model
“Augmenting repository interoperability”

- Repositories as active nodes
- Cross-repository workflow
- Compound objects, e.g. objects within objects + datastreams
- Surrogates

from Herbert Van de Sompel, Southampton workshop on repository interoperability, 4 July 2006

Also to be presented at JISC/CNI Conference, York, 6-7 July 2006
Repository model
Format profiling using PRONOM and ROAR

1. A search is performed in ROAR for archives containing soton (Southampton domain)

2. Clicking the Formats button generates a format summary for all matched archives

3. Clicking a bar shows a breakdown of all files identified as that format (e.g., PostScript 2.0) and associated OAI records

4. Clicking the OAI record identifier shows the Dublin Core record (from which the files were located)

Note: While some PostScript documents also have a PDF version, many do not: this could be the basis for an alerting service to prompt the migration of preservation-unfriendly formats

ROAR http://archives.eprints.org/
Survey of repository policies

Selected repository administrators invited to participate, based on availability and size of ROAR profile

Original test sites for profiling and survey included Oxford University, e-Prints Soton, ECS EPrints (Soton)

Series of questions, based on analysis of preservation metadata for Preserv model

<table>
<thead>
<tr>
<th></th>
<th>EPrints</th>
<th>DSpace</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepted/sent</td>
<td>22</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Returned</td>
<td>13</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Does the repository have any existing policy on preservation?
Does the repository have any existing policy on preservation?

Yes 1  No 18

Example policy
http://www.rub.ruc.dk/rub/selvbetjening/projektbiblioteket_eng.shtml
Does the repository implement any preservation measures, internally or with external agents/services?

<table>
<thead>
<tr>
<th>Measure</th>
<th>Y</th>
<th>N</th>
<th>No reply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte preservation</td>
<td>8</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Transformation</td>
<td>3</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>Rendering</td>
<td>1</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Emulation</td>
<td>0</td>
<td>14</td>
<td>5</td>
</tr>
</tbody>
</table>

Other: backup, mirroring, geographic cluster backup

Partnerships: Sherpa-DP, MetaArchive NDIIPP, dissertation copies at German National Library
Does the repository have a policy on submission file formats?

Y 11 N 4 No reply 4

- prefer PDF / DOC / PPT / HTML
- recommend using PDF or HTML
- PDF (Sherpa policy)
- accept all formats, text documents should be at least be pdf preferred pdf/a
- Use DSpace supported, known, and unknown formats (x3)
- Rendering software must be free, i.e. Acrobat, text, PostScript, HTML
Are there any restrictions (on formats) for submitting authors?

Y 6 N 11 unclear 2

- Word, PDF, Postscript, ascii, html, LaTex, PowerPoint (for conference posters), jpeg (images of book jackets).
- PDF required, other formats optional
- PDF only
- Accept HTML, ASCII, PDF
- ask authors to submit pdf
- majority of deposits PDF; allow other formats, e.g. html, rtf
Does the repository transform submitted formats in any way?

Y 14 N 2 no reply 3

- convert Word docs to PDF
- Most files (e.g. Word, Postscript, PowerPoint) converted to PDF
- transform source files to pdf
- Proprietary formats usually converted to PDF
- transform textual documents to pdf/a
- maths Latex, PS to PDF
- author option to convert to supported format
- video or graphic files zipped
Does the repository require the original source version from the author?

Y 0 N 13 no reply 6

- source file formats (e.g. Word, TeX, WordPerfect) can be deposited
- keep the original sources if deposited
- Authors asked to deposit a copy of "own final corrected draft version" rather than "publisher's formatted version"
Author agreements

Does the repository have any explicit agreement with authors on rights?
Y 16 N 3 ? 0

Examples:
DSpace license with minor changes
"Yes, but mediated deposit means depositing authors don't see it"

Does the agreement refer to rights for preservation?
Y 4 N 10 ? 5
Summary: policy before preservation

Repositories don’t know what they want and are looking for guidance on preservation.

Don’t assume one-size-fits-all service will be sufficient.

Repositories embrace different institutional, cultural and social constraints that will shape policy, including preservation, when they get round to defining it!

Propose a hierarchical series of preservation service models so repositories can choose which one suits.

Repositories are already taking actions on ‘preservation’ that might compromise preservation services.