Preservation as a Process of a Repository

David Tarrant
University of Southampton (UK)
dct05r@ecs.soton.ac.uk
A Few Definitions

**Repository**: A repository is a place where data is stored and maintained.

**IR**: A repository captures and preserves the intellectual output of an institution.

The Case for Institutional Repositories – Raym Crow (SPARC 2002)

**IR**: In my view, 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.

Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age - Clifford A. Lynch

**Service**: A service is something provided directly to a user or 3rd party agent.

David Tarrant, 2008

**Process**: A process is something which is invisible to the user or agent.

David Tarrant, 2008
The Library

• A building to store books in.

• A means by which new books/publications can be acquired.

  • An indexing system to give order.

• Provides a mean by which books can be found.

• Provides a way to borrow & return books.

• A preservation process, e.g. rebind books when they get damaged/worn.

• ...

The Digital Library

The Library

• A building to store books in.
• A means by which new books/publications can be acquired.
• A indexing system to give order.
• Provides a mean by which books can be found.
• Provides a way to borrow & return books.
• A preservation process, e.g. rebind books when they get damaged/worn.
    • ...

The Digital Repository

• A server to store resources on.
• A way to ingest new resources.
• A database of resources and metadata.
• A search engine and dissemination pages.
    • Open access and downloads.
    • A preservation process, e.g. check that the file on the server can still be read/accessed.
    • ...

“In my view, a university-based institutional repository provides a set of services. The repository itself consists of a set of PROCESSES …”
Service: A service is something provided directly to a user or 3rd party agent.

Process: A process is something which is invisible to the user or agent.

There are lots of Processes

Processes happen in parallel

Processes happen in different orders
Many existing models contain this notion of processes and services, just not necessarily in a modern light. This doesn’t however mean they are “wrong” or “right” they are just guiding principals.
Processes in the DCC Model

The DCC Curation Lifecycle Model

CONCEPTUALISE

CREATE OR RECEIVE

TRANSFORM

CURATE

DESCRIPTION

PRESERVATION PLANNING

APPRAISE & SELECT

INGEST

REAPPRaise

STORE

ACCESS, USE & REUSE

MIGRATE

Locate

Retrieve

Store

Index

Ingest

and

Data (Digital Objects or Databases)

and

COMMUNITY WATCH & PARTICIPATION

PRESERVE

PRESERVATION ACTION
The 3 Stage Model

Get Content (Ingest) → Manage Content → Serve Content

Ingest:
- Appraise & Select
- Store
- Index
- Preservation - Check
- Preservation - Analyse
- Preservation - Action
- Dispose

Locate

Retrieve
The manager may provide capability to perform one or more of the processes. Typically the manager is all that is used.
Repository Management Software

A set of Pipes/Workflows* which know how to translate inputs into outputs.

Examples:
- OAI-ORE which contains Files and Metadata is split by the management software into File/Metadata storage and indexes.
- A request for a set of objects related to a single author is translated into a query to an index and a retrieve from the storage.

*Depending on your own definition you could also add “Middleware”
• EPrints Storage Controller works!
  • Local Storage Plugin (legacy)
  • Honeycomb Storage Plugin
  • Amazon Cloudfront (coming soon)

• Honeycomb Stats
  - 4MB/s ingest*
  - 200MB/s retrieve

*USB2.0 max speed
The Preservation Process

Preservation - Check

• Bit checking & checksum calculation

Preservation - Analyse

• What is the type of file, is the file valid?
• Is the file at risk of not having an editor/reader?
• Is there a better format available? Lossless or Lossy?

Preservation - Action

• File migration to avert risks found by analysis.
• Movement of file to new storage.
Preservation - Analysis

- What is the type of file, is the file valid?
  - Droid is a good classification tool for this.

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.

- Is there a better format available? Lossless or Lossy?
  - Planets registry of tools.
What is the type of file, is the file valid?
- Droid is a good classification tool for this.
### Summary Preserv Profile

<table>
<thead>
<tr>
<th>Format</th>
<th>Total Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>[No files found]</td>
<td>9686343</td>
</tr>
<tr>
<td>Portable Document Format (1.4)</td>
<td>267563</td>
</tr>
<tr>
<td>Unknown</td>
<td>257393</td>
</tr>
<tr>
<td>Portable Document Format (1.3)</td>
<td>195221</td>
</tr>
<tr>
<td>Portable Document Format - Archival (1)</td>
<td>180317</td>
</tr>
<tr>
<td>Portable Document Format (1.2)</td>
<td>112050</td>
</tr>
<tr>
<td>Portable Document Format (1.6)</td>
<td>101876</td>
</tr>
<tr>
<td>Hypertext Markup Language</td>
<td>92110</td>
</tr>
<tr>
<td>Portable Document Format (1.5)</td>
<td>82515</td>
</tr>
<tr>
<td>Fixed Width Values Text File</td>
<td>48138</td>
</tr>
<tr>
<td>JPEG File Interchange Format (1.02)</td>
<td>38910</td>
</tr>
<tr>
<td>Tagged Image File Format (3)</td>
<td>32576</td>
</tr>
<tr>
<td>JPEG File Interchange Format (1.01)</td>
<td>30260</td>
</tr>
<tr>
<td>Hypertext Markup Language (4.0)</td>
<td>24476</td>
</tr>
<tr>
<td>Exchangeable Image File Format (Compressed) (2.2)</td>
<td>20572</td>
</tr>
<tr>
<td>Extensible Markup Language (1.0)</td>
<td>18380</td>
</tr>
<tr>
<td>Portable Document Format (1.1)</td>
<td>16780</td>
</tr>
<tr>
<td>OLE2 Compound Document Format</td>
<td>12967</td>
</tr>
<tr>
<td>Extensible Hypertext Markup Language (1.0)</td>
<td>12300</td>
</tr>
</tbody>
</table>
Preservation - Analysis

File Classification

Preserv 2

Formats/Risks

Risks analysis functionality is currently not available. This feature is due to be made available by The National Archives (UK) in the near future. This page will automatically pick up the data when this feature becomes available.

No Risk Scores Available

- Portable Document Format (Version 1.4) [3]
- Microsoft Powerpoint Presentation (Version 97-2002) [3]
- Portable Document Format (Version 1.3) [2]
- ZIP Format [2]
- OLE2 Compound Document Format [1]
Risk Analysis

• Is the file at risk of not having an editor/reader?
  • Functionality is being developed in PRONOM technical registry.

• Simple SOAP web service

• Takes file format identification id’s, hands back risk score.
• Breakdown of risk score may also be available in future releases.

• A stub you can download and run providing this functionality before the official release with mock up risk scores is available at http://preserv2.googlecode.com
This EPrints install is referencing a trial version of the risk analysis service. None of the risk scores are likely to be accurate and thus should not be used as the basis for a program of action.

High Risk Objects

- OLE2 Compound Document Format

Medium Risk Objects

- Microsoft PowerPoint Presentation (Version 97-2002)

Low Risk Objects

- Portable Document Format (Version 1.4)
- Portable Document Format (Version 1.3)
- ZIP Format
### High Risk Objects

<table>
<thead>
<tr>
<th>OLE2 Compound Document Format</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### Medium Risk Objects

<table>
<thead>
<tr>
<th>EPrint ID</th>
<th>File Name</th>
<th>Size</th>
<th>Title</th>
<th>User</th>
<th>No of Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>hitchcock-ipres5-0906-11.ppt</td>
<td>2.64Mb</td>
<td>Towards smart storage for repository preservation services</td>
<td>Mr David C Terrant</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>dordcl2.ppt</td>
<td>11Mb</td>
<td>Applying Open Storage to Institutional Repositories</td>
<td>Mr David C Terrant</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Passig2008_Eprints(97-04).ppt</td>
<td>10Mb</td>
<td>From open storage to smart storage: enabling EPrints repository preservation</td>
<td>Mr Test T User</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EPrint ID</th>
<th>File Name</th>
<th>Size</th>
<th>Title</th>
<th>User</th>
<th>No of Files</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Portable Document Format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Version 1.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Version 1.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Low Risk Objects

<table>
<thead>
<tr>
<th>Portable Document Format (Version 1.4)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Portable Document Format (Version 1.3)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
• Processes, Services and Glue

• Storage Controller provides an API you can glue to.

• Enabling preservation for any repository model by writing small bits of glue.

• Portable services are more powerful, faster and cheaper.

• Make use of existing and supported software where possible.
EPrints will provide one of the first platforms for the development of preservation services where direct interaction takes place between the Repository Software and Preservation Services.
Many Thanks!

David Tarrant
Steve Hitchcock

Neil Jefferies
Ben O’Steen
Sally Rumsey

Adrian Brown
Appendix Slides
Other options for DROID Positioning

This is not the recommended solution as DROID is a 3rd party service for your repository. All other services are provided by your repository.
DROID Alongside Your Resources

- Calendar Server
- Scheduler
- Classification File
- Web Pipe
- Web Server
- Store
  - Smart Storage
    - Preservation - Check
    - DROID
    - OAI-PMH Translator
- Ingest
  - Web Based x 3
  - DROID Classification
- Retrieve
  - OAI-PMH