Connecting Preservation Planning and Plato with Digital Repository Interfaces

David Tarrant
davetaz@ecs.soton.ac.uk
The digital preservation landscape
“75 million people already own iPod Touches and iPhones. That's all people who already know how to use the iPad.”

Picture by curiouslee http://www.flickr.com/photos/curiouslee/4320074421/
“Digital preservation can be simplified by integrating tools and services with existing digital repository interfaces #ipres2010”
EPrints Storage Controller

Bit Preservation

Flexible Decision Logic

Multiple Copies

Multiple Services

Periodic md5 sum checks
# Format Breakdown

<table>
<thead>
<tr>
<th>Format</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrobat PDF 1.4 - Portable Document Format (Version 1.4)</td>
<td>38</td>
</tr>
<tr>
<td>OLE2 Compound Document Format</td>
<td>14</td>
</tr>
<tr>
<td>Acrobat PDF 1.5 - Portable Document Format (Version 1.5)</td>
<td>13</td>
</tr>
<tr>
<td>Acrobat PDF 1.6 - Portable Document Format (Version 1.6)</td>
<td>9</td>
</tr>
<tr>
<td>Acrobat PDF 1.3 - Portable Document Format (Version 1.3)</td>
<td>4</td>
</tr>
<tr>
<td>Microsoft Office Open XML (Version 2007)</td>
<td>2</td>
</tr>
<tr>
<td>Plain Text File</td>
<td>2</td>
</tr>
<tr>
<td>Acrobat PDF/A - Portable Document Format (Version 1)</td>
<td>2</td>
</tr>
<tr>
<td>Windows Portable Executable</td>
<td>1</td>
</tr>
<tr>
<td>Microsoft Word for Windows Document (Version 97-2003)</td>
<td>1</td>
</tr>
<tr>
<td>Acrobat PDF/X - Portable Document Format - Exchange 1:1999</td>
<td>1</td>
</tr>
<tr>
<td>Acrobat PDF 1.7 - Portable Document Format (Version 1.7)</td>
<td>1</td>
</tr>
<tr>
<td>Portable Network Graphics (Version 1.1)</td>
<td>1</td>
</tr>
<tr>
<td>Microsoft Powerpoint Presentation (Version 97-2002)</td>
<td>1</td>
</tr>
</tbody>
</table>
JHove in Plato

Done as part of Preservation Planning

Characterisation

Sample Records

Description of sample records: The sample pages are taken from the 1965, 1971, 1977, and 1983 year books. The quality of these pages is quite similar and so it is for the rest of this part of the collection as it was scanned as one batch. However, the sample images were selected from different years in order to increase the probability for differences in the image quality, as the paper and/or
Risk Analysis in EPrints

Factual vs Subjective Information

Preserv 2

Formats/Risks

- High Risk Objects
  - OLE2 Compound Document Format: 1

- Medium Risk Objects

- Low Risk Objects
  - Portable Document Format (Version 1.4): 3
  - Portable Document Format (Version 1.3): 2
  - ZIP Format: 2

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.
<table>
<thead>
<tr>
<th>High Risk Objects</th>
<th>Graphics Interchange Format (Version 1987a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lada_1200_E.gif (82Kb)</td>
<td>User: Unnamed user with email <a href="mailto:davetz@ecs.soton.ac.uk">davetz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td>No of Files: 9</td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
</tr>
<tr>
<td>Lamborghini_Countach_LP_500.gif (76Kb)</td>
<td>User: Unnamed user with email <a href="mailto:davetz@ecs.soton.ac.uk">davetz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
</tr>
<tr>
<td>Matra_Simca_Rancho.gif (79Kb)</td>
<td>User: Unnamed user with email <a href="mailto:davetz@ecs.soton.ac.uk">davetz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
</tr>
<tr>
<td>Morris_Safran_de_luxe_MK_II.gif (55Kb)</td>
<td>User: Unnamed user with email <a href="mailto:davetz@ecs.soton.ac.uk">davetz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
</tr>
</tbody>
</table>
The Planets Planning Workflow

Preservation Planning in Plato

Define requirements
- Define basis
- Choose records
- Identify requirements

Evaluate alternatives
- Go/No-Go
  - Define alternatives
  - Develop experiment
  - Run experiment
  - Evaluate experiment

Consider results
- Analyse results
  - Set importance factors
  - Transform measured values

Preservation Action Recommendation

Build preservation plan
- Create executable preservation plan
- Define preservation plan
- Validate preservation plan

Planning
- Tree templates and fragments
- Mapping characteristics to requirements

Knowledge base
The Action Plan

A Small part of the Plato preservation plan

```xml
<eprintsPlan>
  <tool>
    <toolIdentifier uri="http://dbpedia.org/data/ImageMagick"/>
    <parameters toolParameters="-verbose -compress None -quality 100 %INFILE% %OUTFILE%"/>
    <targetFormat>PNG</targetFormat>
  </tool>
</eprintsPlan>
```
Performing Actions

- Each set of “at risk” classified files can have a single related preservation plan.
- Once uploaded, any defined actions will be performed on all files of that classification.
Trust, Authenticity and Provenance

Before

Image (GIF)
Download (76Kb) | Preview

After

Image (PNG) (Migrated (Preservation) from Document ID: 41 (image/gif))
Download (76Kb) | Preview
- Image (GIF) (Original Version)
The KeepIT Training Modules

• **Module 1. Organisational issues**
  - Scoping, selection, assessment, institutional parameters

• **Module 2. Costs**
  - Lifecycle costs for managing digital objects, based on the LIFE approach, and institutional costs

• **Module 3. Description**
  - Describing content for preservation: provenance, significant properties and preservation metadata

• **Module 4. Preservation workflow tools**
  - Format management, risk assessment and storage, and linked to the Plato planning tool from Planets

• **Module 5. Trust**
  - The repository’s approach to preservation; trust (by the repository) of the tools and services it chooses
Open Provenance Model (OPM) compliant

Stored in RDF triple form using the EPrints relation manager added in 3.2

OPM @ http://openprovenance.org/
The KeepIT exemplars

Nectar

EdShare

UAL Research Online
“Much more time-consuming and complicated”

“It's not just for the university/ttech people to do it's part of the manager's job too”

“Somewhat less confident…but that's because I know what it really means now”

“I feel well-informed about what is available to help me”

“I would be quite confident teaching other managers, giving a talk etc. to share what I have learned”
“Participation in the KeepIt Project has provided me with an excellent practical grasp of the realities of preservation and brought an apparently enormous and daunting area into the realms of approachable and practical possibilities”
"The KeepIT project has shown me that Digital Preservation is not be the scary thing it can appear to be."

"There isn't a quick fix."

"The secret is to take small steps; to set preservation planning within the bigger institutional context; and to engage stakeholders through advocacy and training"

"My confidence has undoubtedly grown but the more I know, the more I realise there is to know."
“Can we pass XML between systems to enable breakthroughs in digital preservation?”
“Can we pass XML between systems to enable breakthroughs in digital preservation?”