KeepIt

Kultur, eCrystals, EdShare (and NECTAR) – Preserve It!

David Tarrant
davetaz@ecs.soton.ac.uk
School of Electronics & Computer Science
Project Overview

- **Aim:** To create a number of exemplar preservation repositories from which others can learn
- **Small number of very diverse repositories**

Training  Development  Deployment
Repository Content
Preservation

Long Term Reliable Storage

Risk Analysis

Mitigation / Action
Long Term Reliable Storage

- Content specific storage
- Hybrid storage solutions
- Preservation aware storage

- Object orientated storage
- Smart storage
- Simple storage based preservation services

- (Hybrid) EPrints Storage Controller
- EC2 based preservation services (act on S3)
- Storage status reports
Long Term Reliable Storage

EPrints is expanding the number places in which plug-ins can be utilised.

- Import Plug-ins
- Export Plug-ins
- EPrints Core Interfaces, Submission Manager
- Database Controller
- Storage Controller

CLOUD (Amazon S3)
Risk Analysis

- File Formats
- Significant Properties
- Resolution policy

- Risk analysis policies
- Preservation registry (p2-registry)

- File format plug-in toolset EPrints
- Risk analysis for files in repository
- An open preservation registry
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
- Portable Document Format (Version 1.3): 2
- ZIP Format: 2
- OLE2 Compound Document Format: 1
Risk Analysis - Portable Document Format (v1.3) (Default Profile)

Portable Document Format (v1.3)

- Format Age: Your format is 10 years old and there are 3 newer formats, the latest of which is PDF (1.6) (Released: 01 Jan 2004).
- Software Tools (Open): 3 tools can Open your format.
- Software Tools (Save): 1 tool can Save your format.
- Format Documentation: Documentation exists for this format.
- Documentation Quality: Documentation is complete and of a high standard.
- Rights: Format is proprietary.

Portable Document Format

- Format Age: Your format is 16 years old but is the latest known version of this format.
- Ubiquity: Format is most widely adopted of type.
- Stability: Format is not backwards compatible, but versions change infrequently.
- Identification Type: Format can be positively identified (specific).
- Format Type: It is not possible to obtain the original document in the original context using this format.
- Complexity: Medium complexity format.
- Software Tools (Open): 14 tools can Open your format.
- Software Tools (Save): 39 tools can Save your format.

Risk Score: 3.73

How is this calculated?

The data you see here has all come from the Preserv2 registry and more specifically the risk analysis service. Available here in RDF the risk analysis service takes specific information from the registry according to a profile (in this case the default one) and outputs in in RDF. This page displays a summary of this data which has also been processed to find a score relating to this data.

Each piece of select data is either about the format itself or its related supertype format, e.g. PDF 1.6 is a type of PDF. From this point data is handled in 4 ways with all final risk levels being either Low (<3.5), Medium (3.5 - 6.99), or High (>= 7).

To calculate the final risk score low risks are worth 1 point, medium - 5 points and high - 10 points. The total is then divided by the number of properties which counted towards this score to give the final risk score. Items with lines through them are not counted due to better or more accurate overriding information being available in a different category.

The risk boundaries are:

- <3.51 = Low Risk
- 3.50 and <7.00 = Medium Risk
- >=7.00 = High Risk
Risk Analysis

Risk analysis for files in repository

Preserv 2

Formats/Risks

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
Mitigation / Action

- Choosing storage platforms
- Migration vs Emulation
- The importance of policy

- Migration pathways
- Integration of simple migration services
- Using smart storage and cloud based services

- Risk Analysis -> Migration Integration
## High Risk Objects

<table>
<thead>
<tr>
<th>Tool</th>
<th>Preservation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPT -&gt; PPTX</td>
<td>Blue</td>
</tr>
<tr>
<td>PPT -&gt; PDF</td>
<td>Green</td>
</tr>
</tbody>
</table>

## Medium Risk Objects

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Title</th>
<th>EPrint ID</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>hitchcock-ipres5-0908-11.ppt</td>
<td>2640Kb</td>
<td>Towards smart storage for repository presentation services</td>
<td><strong>4</strong></td>
<td>Mr David C Tarrant</td>
</tr>
<tr>
<td>dorsdl2.ppt</td>
<td>11Mb</td>
<td>Applying Open Storage to Institutional Repositories</td>
<td><strong>1</strong></td>
<td>Mr David C Tarrant</td>
</tr>
<tr>
<td>Passio2008_Eprints(97-04).ppt</td>
<td>10Mb</td>
<td>From open storage to smart storage: enabling EPrints repository preservation</td>
<td><strong>2</strong></td>
<td>Mr Test T User</td>
</tr>
</tbody>
</table>

## Low Risk Objects

<table>
<thead>
<tr>
<th>Tool</th>
<th>Preservation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portable Document Format (Version 1.4)</td>
<td>Green</td>
</tr>
<tr>
<td>Portable Document Format (Version 1.3)</td>
<td>Green</td>
</tr>
</tbody>
</table>