A preservation system

Challenges & practical experience

Digitálna Knižnica | Jasná, Slovakia | March 2009

Yaniv Levi, Rosetta product manager
Copyright Statement

All of the information and material inclusive of text, images, logos, product names is either the property of, or used with permission by Ex Libris Ltd. The information may not be distributed, modified, displayed, reproduced – in whole or in part – without the prior written permission of Ex Libris Ltd.

TRADEMARKS
Ex Libris, the Ex Libris logo, Aleph, SFX, SFXIT, MetaLib, DigiTool, Verde, Primo, Voyager, MetaSearch, MetaIndex and other Ex Libris products and services referenced herein are trademarks of Ex Libris, and may be registered in certain jurisdictions. All other product names, company names, marks and logos referenced may be trademarks of their respective owners.

DISCLAIMER
The information contained in this document is compiled from various sources and provided on an "AS IS" basis for general information purposes only without any representations, conditions or warranties whether express or implied, including any implied warranties of satisfactory quality, completeness, accuracy or fitness for a particular purpose.

Ex Libris, its subsidiaries and related corporations ("Ex Libris Group") disclaim any and all liability for all use of this information, including losses, damages, claims or expenses any person may incur as a result of the use of this information, even if advised of the possibility of such loss or damage.

© Ex Libris Ltd., 2008
Short background

- ExLibris has built a digital preservation system Rosetta with the National Library of New Zealand.
- Version 1.0 of the Rosetta was released in January 2009.
- The system includes ingest, validation, delivery and a permanent repository.
What was involved

• 23 specs with over 1600 pages
• 9 Modules covering the preservation workflow
• A team of over 25 FTE’s
• 2 Years of development work
• System released – Dec 08’
• Version 2.0 will be released for beta – Dec 09’
What is digital preservation?

Or how to deal with the jungle of definitions and understanding
The Paradox of digital preservation

“Traditionally, preserving things meant keeping them unchanged; however our digital environment has fundamentally changed our concept of preservation requirements. If we hold on to digital information without modifications, accessing the information will become increasingly difficult, if not impossible”

Su-Sing Chen, Computer, March 2001, 2-6
Preservation - definition

Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.
Preserving
(Active preservation)

Collecting

Archiving
(Passive preservation)
Challenge II

The data model...

Or how to build a best practice when none are in existence...
• **Goal:**
  
  • Need support for long term digital preservation.
  • Data model to support a variety of materials and fields.
  • Needs to be standard and open

• **Solution:**
  
  • Data model based on PERMIS and METS
  • Flexible structure and allow extensions to PERMIS
  • Support for ISAD(G) and archival concepts
Intellectual Entity

- Representation Master
  - JP2
  - JP2
  - JP2
  - JP2

- Representation Original
  - TIFF
  - TIFF
  - TIFF
  - TIFF

- Representation Access copy
  - PDF
Challenge III

DROID, JHOVE, PRONOM and other open source tools

Or how to manage a set of tools that not necessarily work together with no UI interfaces and have not been updated for a while.
• **Goal:**

You need to perform characterization to know what you have, and the quality of the information.

• **Solution**

Use a set of open source tools such as DROID, JHOVE and New Zealand extracting tool to provide characterization solution.

• **Problems:**

  • Overlapping functionality in the tools
  • Scalability issues
    • E.g. DROID has 50% failure
    • 18 sec. for validation.
  • No UI to manage the tools or exceptions
Solution in Rosetta

- A framework around the different tools including:
  - A technical analyst workbench to handle the different issues.
  - Rule based decision making to automate the process.
  - Scalable infrastructure.
  - Use the local format library to define rules of what to use on each format.
### SIPs List

<table>
<thead>
<tr>
<th>SIP ID</th>
<th>Deposit ID</th>
<th>Title</th>
<th>Producer</th>
<th>Type</th>
<th>No. of Files</th>
<th>Problems</th>
<th>Assigned To</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4 Problematic files</td>
<td>Demo Internal Producer</td>
<td>UNP</td>
<td>5</td>
<td>50% failed</td>
<td>admin1</td>
<td>Work on</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6 Title</td>
<td>Demo Internal Producer</td>
<td>UNP</td>
<td>5</td>
<td>33% failed</td>
<td>admin1</td>
<td>Work on</td>
</tr>
<tr>
<td>3</td>
<td>143</td>
<td>183 Aliza test</td>
<td>Photograph Association</td>
<td>UNP</td>
<td>2</td>
<td>100% failed</td>
<td>admin1</td>
<td>Work on</td>
</tr>
<tr>
<td>4</td>
<td>145</td>
<td>185 Commit1</td>
<td>Photograph Association</td>
<td>UNP</td>
<td>5</td>
<td>40% failed</td>
<td>admin1</td>
<td>Work on</td>
</tr>
<tr>
<td>5</td>
<td>162</td>
<td>202 Problems 2</td>
<td>Photograph Association</td>
<td>UNP</td>
<td>5</td>
<td>60% failed</td>
<td>Demo5</td>
<td>View</td>
</tr>
</tbody>
</table>

**Actions**
- Move To Next Step
- Execute

**Technical Details**
- Virus Check: 0 files
- Fidelity Check: 0 files
- Format Validation: 3 files
- Technical MD Extract: 1 file
<table>
<thead>
<tr>
<th>File Name</th>
<th>PID</th>
<th>Status</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format.dat</td>
<td>FL2010</td>
<td></td>
<td>1 Tasks failed, Download, Replace, Recheck, More</td>
</tr>
<tr>
<td></td>
<td>format.unknown</td>
<td>FL2011</td>
<td>2 Tasks failed, Download, Replace, Recheck, More</td>
</tr>
<tr>
<td></td>
<td>file_600.eig</td>
<td>FL2012</td>
<td>1 Tasks failed, Download, Replace, Recheck, More</td>
</tr>
</tbody>
</table>

**Fixity Check:** Pass
File Format Fail - Tentative format identification (vs_Error 9)
Virus Check: Pass
<table>
<thead>
<tr>
<th>Active</th>
<th>Set Order</th>
<th>Name</th>
<th>Description</th>
<th>Created Date</th>
<th>Updated Date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Multiple Text Files file formats</td>
<td>Define a single definite file format for text files</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Tentative MP3 Files file format</td>
<td>Define the file format of MP3 files as definite an...</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Multiple Excel Files file formats</td>
<td>Define a single definite file format for Excel files</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Multiple Tiff Files file formats</td>
<td>Define a single definite file format for Tiff files</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Multiple RTF Files file formats</td>
<td>Define a single definite file format for RTF files</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Multi Word doc</td>
<td>Matches DIFF and MS word, match on ext to make sur...</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Tentative Jpeg2 Files</td>
<td>Map tentative Jpeg2000 files</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Tentative LOG files</td>
<td>Map tentative LOG files</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Tentative ARC file format</td>
<td>ARC format correction rule</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Tentative PDF file format</td>
<td>PDF format correction rule</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Multiple RTF v1 5</td>
<td>Define a single definite file format for v1.5 RTF ...</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Multiple RTF v1 7</td>
<td>Define a single definite file format for v1.7 RTF ...</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Catch anything else</td>
<td>This is for migration purposes only (must b...</td>
<td>16/12/2008 11:43:23</td>
<td></td>
<td>Update</td>
</tr>
</tbody>
</table>
PRONOM

- PRONOM is a global format registry

- **Problem:**
  
  Is not being updated very often.

- **Solution:**

  build a local format library that will add information to the one in PRONOM and will know how to work with PRONOM side by side.
Challenge IV

Integrating into the library existing infrastructure

Or how not to invent yet another wheel with similar functionality that already exists in the library.
• Create an open system that can use the infrastructure in the library.
  • Integration with Descriptive tools and applications.
  • Submission applications (SDK).
  • Integration with IAM.
  • Patron search and display tools.
Integration Points

Customer systems uses

- Validation Plug-in
- Metadata Extraction
- Enrichment Plug-in
- Persistent Identifier
- CMS Integrations
- Access API
- Viewers
- Web Services
- Publishing
- Migration Tools

Digital Preservation

- Deposit
- Working Area
- Operational
- Permanent
Challenge V

Assuring that the system is not a preservation bottleneck by itself

Or how do we ensure that the information will not be locked into the system forever.
• All preservation worthy information is stored on disk which is a proven technology.
• The storage on disk is open and based on standards (PERMIS, METS etc.)
• The information can be restored into Rosetta or into another system.
Challenge V

Conform to OAIS

Or OAIS is just a reference model how do you conform to it?
OAIS and Rosetta

OAIS

- Ingest
- Archival Storage
- Data Management
- Access
- Preservation Planning

Rosetta

- Deposit
- Working Area
- Permanent
- Operational Management
- Publishing
- Delivery
- Search Tools
- Preservation
Thank You!

Yaniv.Levi@ExLibrisGroup.com