Rosetta
A digital preservation solution for Universities

Josh Weisman
Development Director, Ex Libris Ltd.
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., 2012
What is Rosetta?

Rosetta is a **complete** preservation solution that addresses the **ever-growing** need to collect, archive and preserve the **digitally-born and digitized** materials stored at academic institutions, research organizations, and government institutions, ensuring **data integrity** and **access over time** to information in digital formats.
Agenda

1. The need for digital preservation
2. The challenges
3. The Rosetta Architecture
4. Who is using Rosetta and how
5. Longterm preservation for Universities
The need for digital preservation
Need for Digital Preservation

Today’s world is digital. If a file can’t be opened, probably the reasons are:

1. Corrupted media
2. Missing rendering application
3. Unidentified file format
Need for Digital Preservation

Nothing digital lasts:

1. Media has a shelf life
2. Applications have a shelf life
3. Formats have a shelf life

You need to preserve all three to be able to read it again
The Challenges
Challenges

Active preservation principles:

1) Ensuring bit integrity

2) Ensuring content health
   • Format viability
   • Complete metadata
   • Provenance

3) OAIS compliant system
Challenges - Bit Integrity

- Fixity checks determine if data has changed or corrupted
- Basic feature found in asset management as well as preservation solutions
- Does not guarantee data access – just that it has not changed

A small first step to preservation
Challenges – Content Health

- Formats evolve rapidly and become obsolete
- File access requirements
  - Positive ID of format e.g. pdf
  - SW application e.g. Acrobat reader
- Risk analysis to assure access
  - Current format library
  - Current application library

Essential for preservation
Challenges – Complete metadata

• Technical metadata (e.g. size, resolution, compression, etc)

• Descriptive metadata (e.g. author, title, publisher, etc)

• Metadata is essential so that you
  • Preserve exactly what you need to preserve
  • Know (in future) what you preserved
Challenges - System

A digital preservation system (OAIS) must offer ¹

¹. International Organization for Standardization: Open Archival Information System (OAIS)
OAIS Model
The Rosetta Architecture
Rosetta Architecture
Key Features

- Community Driven Knowledge Base
- Active Preservation
- Scalable
- Open & Integrative
- Digital Asset Management
- Out-of-the-box Configuration
Rosetta - Open & Integrative

Submission Apps

Storage Abstraction

Rosetta

Plug Ins (validation, migration, enrichment, etc)

Search Engines

ILS/CMS Systems
Rosetta as Digital Asset Management solution

- Rosetta consists of several DAM-features (e.g. collection management)
- This enables customers to:
  - Replace existing systems
  - Use Rosetta as repository for example in digitization projects
Who’s Using Rosetta and How
Rosetta Customers Around the World
## Different focus areas

<table>
<thead>
<tr>
<th>National / State libraries</th>
<th>Archives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Cultural heritage</td>
<td>• Legal documents</td>
</tr>
<tr>
<td>• Websites</td>
<td>• Archival collections</td>
</tr>
<tr>
<td>• Mass digitization</td>
<td>• Government papers</td>
</tr>
<tr>
<td>• Maps</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Special libraries</th>
<th>Universities &amp; Research Institutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Audio-Video files</td>
<td>• Research data</td>
</tr>
<tr>
<td>• Special collections</td>
<td>• Digitized student files</td>
</tr>
<tr>
<td>• Publications</td>
<td>• Certificates</td>
</tr>
<tr>
<td>• E-mails</td>
<td>• DICOM-files</td>
</tr>
<tr>
<td></td>
<td>• Administrative folders</td>
</tr>
</tbody>
</table>
Growing need for preserving digital information in Universities
Why?

• More and more information in University organisations (incl. University-Hospitals) has been digitized or is born-digitally

• Examples
  • In the Library: publications, special collections, images, maps, audio-video files, etc.
  • In the Administration: Digitized Certificates (need to be accessible by law for 40 years), digitized student files, etc.
  • In the Hospital: Electronic health files, X-ray files (DICOM format)
  • In the faculties: Research data
  • In the University Archive
  • More
Common requirements

These institutions must collect, manage and preserve their content and provide long term access to this information.
Focusing on the area of research data
Why should Universities care?

- Good scientific practice requires retention of data in usable form
- Research collaboration
- Funding organisations require data management plans (NSF, DFG)
- Re-use of data increasingly important and should be facilitated
- Data which cannot easily be reproduced and has permanent relevance must remain available
- Published or referenced material must be citable and remain available
**Funders’ View – The UK example**

<table>
<thead>
<tr>
<th>Research Funders</th>
<th>Policy Coverage</th>
<th>Policy Specifications</th>
<th>Support Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Published outputs</td>
<td>Data</td>
<td>Time limits</td>
</tr>
<tr>
<td>AHRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBSRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRUK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPSRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NERC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STFC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
Challenges

- **Missing data management process/Who is responsible?**
  - access rights control
  - retention periods
  - DOI
  - National initiative? University/Library?

- **Meta data definition**
  - Heterogenous disciplines
  - Standardisation?

- **Legal aspects eg. Copyright/intellectual property**

- Get the researchers on-board
Example: ETH Solution

Data production and handling for current analysis

Pre-ingest, e.g. structuring, re-arranging, selecting

Long-term preservation according to OAIS

Additional search/retrieve & delivery components

Access components

Catalog
Admin-Interface
Archiv-DB
Repositories

Create Deliveries / Retrieve Requests

User

Additional Ingest components

Additional Ingest

Security-Layer (Authentication, Authorisation)

Storage-Layer

Content sources

Access components

Security-Layer

Producer

ETH - Pre-ingest Mechanism
Duesseldorf Proof-of-Concept

Goals:
• Demonstrate ingest capabilities of the four different data types including format validation and identification
• Demonstrate preservation capabilities
• Demonstrate integration with external viewers
• Demonstrate publishing of the research data

Data Types:
• Biological and medical research center (BMFZ, genetics data)
• Center of advanced imaging (CAi, microscopic data)
• University library (ULB, METS/TIFF data)
• University hospital (UKD, DICOM data)

Results:
All items are stored in the permanent repository, some can be viewed and published.
Where are we today?

Using Rosetta you can:

1. Allow researchers to deposit research data materials both ‘light’ deposits like Excel files and ‘heavy’ deposits like DNA sequence file (40gb)
2. Assign different retention policies
3. Plugin your own viewers or export the materials to dedicated machines
4. Publish materials using Primo or other resource discovery tools
5. Create DOIs

• And more...
Recommended solution concept for University data centers
Benefits of this solution concept

• The University is able to archive and preserve all digital material of the entire organization in one system.

• For some units of the university Rosetta can be used as a repository which will reduce maintenance and operation effort for the University IT.

• Long-term preservation can be offered as a new service for the university which is especially of interest to the faculties producing plenty of digital research data.

• The pre-ingest tool helps researchers organize and manage their data before it is ready for archive.

• Partners of the university might join in the project to reduce effort / costs for operation and maintenance.
Thank You!