LIBER QUARTERLY 21 (1), November 2011 — ISSN: 1435-5205. P138–145
http://liber.library.uu.nl/
Igitur publishing

The Challenges of Building Ex Libris Rosetta, a Digital Preservation System

Ido Peled

Ex Libris,
ido.peled@exlibrisgroup.com

Abstract

This article describes challenges regarding digital preservation that arose during the development of the Ex Libris Rosetta digital preservation system.

Key Words: digital preservation; digital preservation system

Introduction

In the last two decades, digital technology has enabled us to create, use, and be enriched by information in ways that were unthinkable a generation ago. The growth in the number of digital items in today's library collections — items that have undergone digitization and items that were 'born' digital — has led to an understanding that new actions must be taken to preserve these digital assets and make them available to future generations.

The challenge of preserving digital material is most acute regarding items that were 'born' digital. The vast majority of this material exists exclusively in digital format, a fact that makes the preservation of digital information critical to the perpetuation of our cumulative knowledge.

While many organizations have systems in place for storing and managing digital objects, these systems are not always designed with preservation in mind. Digital preservation is about guaranteeing the continued usability of and access to digital content tomorrow and well into the future. Digitalasset management systems and digital repositories focus on facilitating the day-to-day use of digital content, whereas a digital preservation system offers discovery and access options, functionality and workflows for ingesting materials, ongoing risk analysis, and the continual integrity of stored items. Although preservation focuses on risk management, we would be mistaken if we equated preservation with back-up or disaster recovery.

The Open Archival Information System (OAIS) reference model describes the characteristics of a digital preservation system.[1] The model has become widely accepted among preservation bodies and experts worldwide and has been used as a guideline to evaluate current implementations of preservation and archiving initiatives.[2] The OAIS model specifies six high-level functions that must be present in a digital preservation system:

- Ingest
- Storage
- Data management
- Administration
- Preservation planning
- Access.

Moreover, the OAIS reference model helps us understand the workflows required for any preservation system as well as outlining the terminology for digital preservation. Terms such as SIP (submission information package), AIP (archival information package), DIP (dissemination information package), producer, and customer should be part of every preservation-minded person's vocabulary.

The six functions described in the OAIS model are, indeed, an integral part of Ex Libris Rosetta, which was released in January 2009. Developed in partnership with the National Library of New Zealand and reviewed by a peer group of world-renowned preservation experts, Ex Libris Rosetta addresses the need of libraries, archives, and other memory institutions to collect, manage, and preserve a wide variety of digital objects in different formats and structures.

Ex Libris Rosetta enables institutions to manage digital entities end to end — from submission to dissemination. A rule-based workflow engine and open architecture allow institutions using the system to develop unique plug-in tools and other applications to enhance the system's ingest, management, preservation, and delivery processes.

This article describes challenges regarding digital preservation that arose during the development of the Ex Libris Rosetta digital preservation system.

Challenge I: The Meaning of Digital Preservation, Or how to deal with the jungle of definitions

Defining the end goal of any digital preservation system is easy: ensuring the integrity of the system's digital content and access to it over time. However, digital preservation itself can be defined in various ways, which usually include a specification of what must be done in order to preserve digital content, including the following tasks:

- **Disaster recovery:** Identifying the risks of disaster and ways of recovering from those risks applies to the preservation not only of digital content but also of physical paper.
- **Back-up:** Ensuring that content is backed up and that several copies are stored (ideally not in the same location) makes recovery from disaster possible and also maintains the files' bit-level integrity.
- **Media refreshment:** The storage medium holding the various backups has a limited life and carries other types of risks as well, such as loss of data authenticity and integrity, data destruction, and data degradation. In a recent study by the PrestoPRIME consortium (a project funded by the European Union for the preservation of audiovisual content), more than 35 storage risks were noted.[3] Through the process of media refreshment, institutions can keep moving their content to new storage media while evaluating the risks and alternatives of alternative formats.
- **Preservation planning:** Preservation planning services empower organizations to define, evaluate, and execute preservation actions, such as format migration and the implementation of emulators. By using preservation planning functions and actively refreshing the storage media, organizations keep their digital materials accessible.

The Association for Library Collections & Technical Services (ALCTS) has several definitions for digital preservation, one of which is as follows: '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.'[4] This definition identifies that digital preservation is a process rather than a snapshot and also refers to different types and origins of materials — digital-born and digitized.

Ex Libris identifies three main topics that should be included in a definition of digital preservation:

- **Collecting:** collecting the content from producers and facilitating the deposit of digital materials into the system
- **Archiving:** ensuring the bit integrity of the files
- **Actively preserving:** actively assessing risks on the files in the repository, constantly evaluating ways to mitigate the risks, and executing actions to resolve the risks.

Conforming to the Open Archival Information System (OAIS) standard, Ex Libris Rosetta provides preservation planning and preservation actions that enable institutions to manage objects in multiple digital formats, detect and mitigate format-related risks, and perform a variety of preservation-related tasks.

Challenge II: Data Model, Or how to build a best practice

Creating a data model for a digital preservation system consists of several tasks and challenges:

- Support a structure and metadata for long-term digital preservation
- Support a variety of formats and materials in one data model
- Support a variety of institution types (museums, libraries, and archives) with preservation needs
- Conform to standard and open practices.

Rosetta's data model is based on the Preservation Metadata: Implementation Strategies (PREMIS)[5] working group's metadata elements and the Metadata Encoding and Transmission Standard (METS)[6]:

- PREMIS as the conceptual model outlining the entities and metadata required for preservation
- METS as the wrapper for encoding descriptive, administrative, and structural metadata.

A preservation system must also document and preserve the descriptive information of its digital entities, but it is by no means a cataloguing system. For that reason, Rosetta uses the extendable Dublin Core elements to store descriptive metadata in the system.

Challenge III: Characterization, Or how to identify and validate content by means of the most up-to-date tools

Characterization means knowing what you have and knowing that what you have is viable and valid. Knowing what you have is the first step of preservation; moreover, no preservation action can take place without proper technical information about the files in the repository.

The challenge of characterization includes understanding what is held in the repository (what formats, which versions, and so on), ensuring that the content is not corrupted, and storing the information for future use.

Rosetta answers these challenges through its Working Area module, which uses the Digital Record Object Identification (DROID)[7] software tool (developed by the National Archives in the UK) to perform automated batch identification of file formats. Rosetta extends the file identification by allowing a rule-based decision-making configuration to determine formats that DROID does not identify.

The characterization and extraction of technical metadata is carried out by JSTOR/Harvard Object Validation Environment (JHOVE)[8] and the National Library of New Zealand Metadata Extraction Tool[9]. Rosetta's open architecture also allows other third-party plug-in tools to be used for extracting metadata from a file. To ensure that the file was ingested into the system intact, the file is checked for viruses and a fixity check is performed.

To make better use of these tools and allow the integration of future characterization and validation tools, Ex Libris developed a highly scalable framework that uses rule-based decision-making software. In addition, Rosetta contains a workbench designed to deal with the various issues that may arise during the characterization process.

Challenge IV: Existing Infrastructure, Or how not to invent yet another wheel with similar functionality

Most institutions already have an infrastructure in place, whether it is used for digital collections or for cataloguing paper materials. The challenge when building a digital preservation system is triple: first, the system should not attempt to replace all existing repositories and tools in use by the institution. Second, the system should be built to be used for preservation first and not last. Third, the developers should ensure that the system can be integrated with the existing infrastructure and can communicate with the existing environment.

One of the guiding principles when building Rosetta was that it should follow the Ex Libris open architecture policy. The goal was to create a software development kit (SDK) for Rosetta, along with the many Web services and application programming interfaces (APIs). By using these tools, institutions can develop complete applications for ingest, discovery, and more. In addition, the system now includes a plug-in manager that facilitates the use of third-party tools.

Challenge V: The System as a Bottleneck, Or how to ensure that information will not be locked in the system forever

Like other software, a digital preservation system is built in a specific environment and with specific tools and software. The challenge that we faced when designing Rosetta was to ensure that the system itself will not be a preservation bottleneck — in other words, that proprietary Ex Libris elements will not lock in the content forever. Even with the most current technology and tools, there is no way to avoid obsolescence in a few decades at best, or in a few years at worst. From the start, a digital preservation system must include an 'exit' strategy to ensure the longevity of the digital content well beyond the system software's lifetime.

To meet this challenge, the Rosetta permanent storage area was created as the final destination of the files, where they are archived after validation, enrichment, and manual assessment. The permanent storage area holds not only the content files but also an open, easily read and understood XML document containing all the information about the file, including the descriptive and technical metadata, the file's access rights and provenance events. In fact, the permanent storage area holds a complete replication of all the data used by the system.

The replicated data provides a way for institutions to retain control over their data; should Rosetta become obsolete, institutions will be able to harvest their data from the storage area and reconstruct the entire repository of the preservation system.

The replication of stored data also acts as a redundant layer that will make a complete restoration of the system possible if a disaster affects the system or its database.

Conclusions

The main challenges of building a preservation system include the following:

- Defining digital preservation
- Creating an appropriate data model
- Characterizing and identifying the digital content
- Integrating the system with the existing infrastructure
- Ensuring the longevity of the data and the system

Although these challenges and their solutions were critical when we designed the Rosetta digital preservation system, they represent only a small proportion of the requirements and objectives that we defined for the system. Technology, changing standards, and customer expectations and requests introduced a set of new challenges that needed to be met.

Now in its second major version (version 2.1 was released in November 2010), Ex Libris Rosetta is used at numerous institutions around the world to collect, archive, and preserve their digital collections.

Confidential information — Disclaimer

The information herein is the property of Ex Libris Ltd. or its affiliates and any misuse or abuse will result in economic loss. Do not copy unless you have been given specific written authorization from Ex Libris Ltd. This document is provided for limited and restricted purposes in accordance with a binding contract with Ex Libris Ltd. or an affiliate. The information herein includes trade secrets and is confidential.

The information in this document will be subject to periodic change and updating. Please confirm that you have the most current documentation. There are no warranties of any kind, express or implied, provided in this documentation, other than those expressly agreed upon in the applicable Ex Libris contract.

Any references in this document to non-Ex Libris web sites are provided for convenience only and do not in any manner serve as an endorsement of those web sites. The materials at those web sites are not part of the materials for this Ex Libris product and Ex Libris has no liability for materials on those Web sites.

Notes

[1] http://public.ccsds.org/publications/archive/650x0b1.pdf.

[2] See, for example, the *Assessment of UKDA and TNA Compliance with OAIS and METS Standards* report, at http://www.jisc.ac.uk/uploaded_documents/oaismets.pdf.

[3] http://prestoprimews.ina.fr/workpackages/wp3_data/deliverables/release/PP_WP3_ID3.2.1_ThreatsMassStorage_R0_v1.00.pdf/view.

[4] http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.cfm.

[5] http://www.loc.gov/standards/premis/.

[6] http://www.loc.gov/standards/mets/.

[7] http://droid.sourceforge.net/.

[8] http://hul.harvard.edu/jhove/.

[9] http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool.