Ex Libris Rosetta:

A Digital Preservation System

Product Description

CONFIDENTIAL INFORMATION

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.

DISCLAIMER

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.

Copyright Ex Libris Group, 2010. All rights reserved.

Documentation produced February 2010.

Web address: http://www.exlibrisgroup.com

Index

Executive Summary

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 created in digital format—that is, "born" digitally. 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. Digital asset 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 backup or disaster recovery.

The Open Archival Information System (OAIS) reference model describes 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:

---

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

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

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

These six functions, indeed, are an integral part of Ex Libris Rosetta for digital libraries, whose release was announced in January 2009. Developed in partnership with the National Library of New Zealand and reviewed by a peer review group of world-renowned preservation experts, Ex Libris Rosetta addresses libraries' and archives' need to collect, manage, and preserve a wide variety of digital objects in different formats and structures.

Ex Libris Rosetta enables libraries 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 applications to enhance the system's ingest, management, preservation, and delivery processes.

Product Information

High-Level Architecture

Ex Libris Rosetta is a digital-object preservation solution that conforms to the ISO-recognized Open Archival Information System (OAIS) and supports international industry standards such as the Metadata Coding and Transmission Standard (METS), Preservation Metadata: Implementation Strategies (PREMIS), Dublin Core, and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The system is designed to support the acquisition, validation, ingest, storage, management, preservation, and dissemination of different types of digital objects and adheres to e-legal deposit requirements.

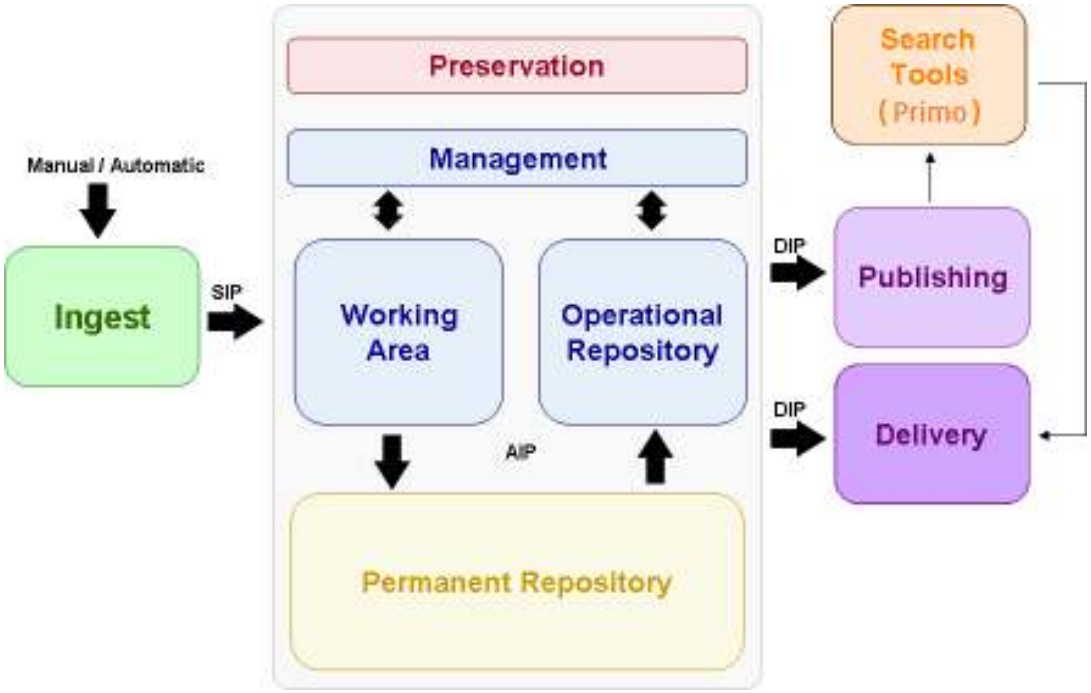

Figure 1. Ex Libris Rosetta system architecture

Ex Libris Rosetta consists of a set of separate yet interactive modules that correspond with the life cycle of a digital object.

Deposit

The deposit module enables external producers (publishers) or internal producers (institution staff) to upload and save submission information packages (SIPs)[3] for ingestion into the preservation repository. The module supports multiple workflows—automatic, semiautomatic, and manual—which can be assigned to the producer depending on the producer's agreement with the institution. The deposit module provides producers with a Web-based user interface through which they can upload files, provide metadata, and define access restrictions. Files and metadata can also be uploaded to an FTP site from which the deposit module downloads them. A software development kit (SDK) provides seamless integration of the deposit module with record-management systems, thus facilitating the process of making large, frequent deposits.

A SIP can contain one or more intellectual entities (IE).[4] An intellectual entity can be either simple (a single digital entity or file to be managed and preserved) or complex (multiple digital entities or files that are managed and preserved as one or more groups).

[3]As explained in the draft *Submission Information Package (SIP) Specification* version 1.0, "under the OAIS model, material from a content provider is transmitted to the archive in a form called a Submission Information Package (SIP). The archive Ingest function accepts the SIP and potentially transforms its contents into an internal form called an Archival Information Package (AIP) for long-term preservation".

[4]As defined in the PREMIS Data Dictionary for Preservation Metadata, an intellectual entity is "a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more digital representations" (http://www.loc.gov/premis/v2/premis-2-0.pdf).

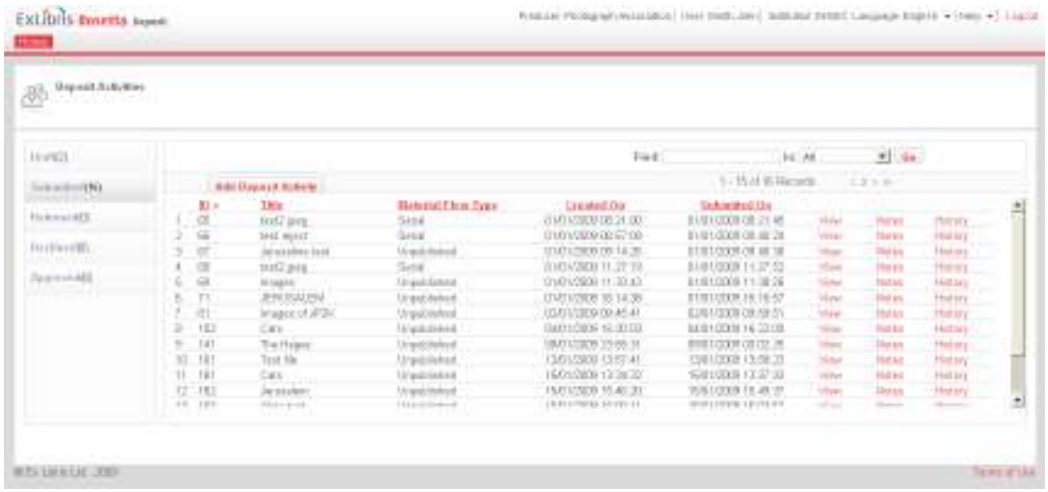

Figure 2. Deposit area with a list of submitted items

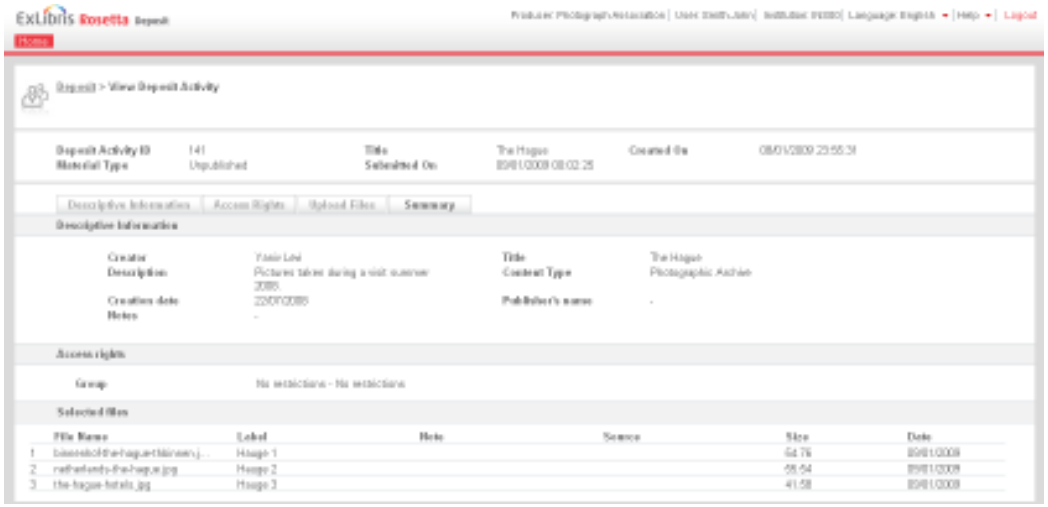

Figure 3. Summary of a deposit activity in the deposit module

Working Area

The working area is a repository for managing internal ingests processes for materials that have been submitted to the system. It represents the point at which a SIP becomes an archival information package (AIP) in the OAIS model.

The working area supports both automatic and manual processes, such as the following:

- Virus checks, checksums, and format identification and validation to ensure the integrity of the SIP data stream (files) and metadata
- Curatorial and accession processes that monitor the quality of the submitted material to verify that it is complete and not damaged

- Enrichment processes that convert information into a stable format when necessary and enrich objects with relevant metadata

Some of the actual processes (for example, virus checks and format validation) are carried out by means of third-party tools, such as designated user-interface tools and APIs. The working area's workflow engine facilitates the embedding of such tools.

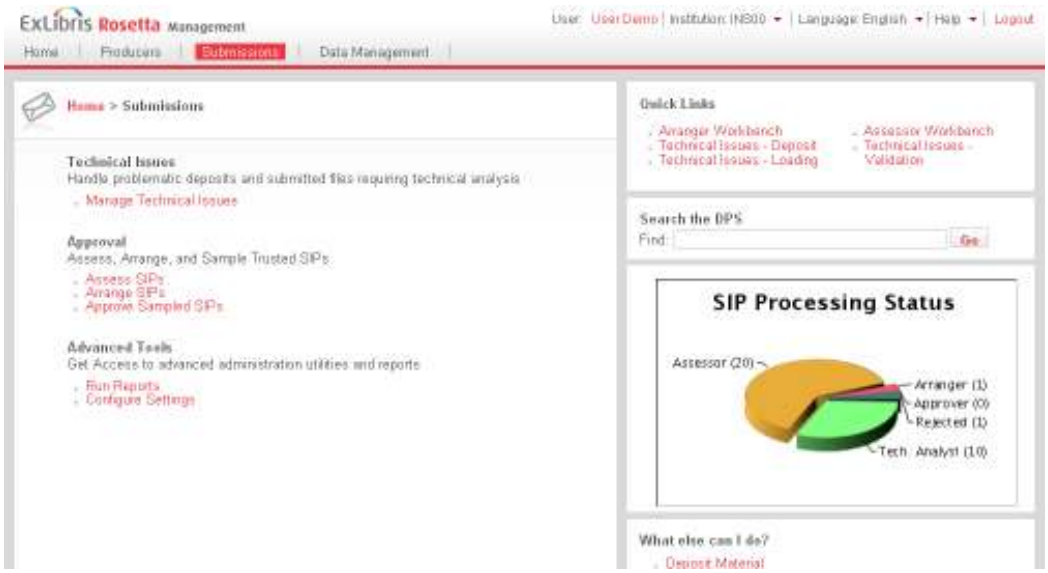

Figure 4. Submissions in the working area

Permanent Repository

The permanent repository module is a long-term preservation repository for approved and successfully enriched materials. Its core function is to store information in perpetuity. Items are stored under a write-once, read-only policy with full disk replication to ensure that data and objects are completely secure.

Full disk replication provides the following benefits:

- The system's dependence on the database is minimal, because everything can be reconstructed from the repository disk.
- Information about each object is self-contained and includes full, relevant metadata for managing preservation actions over time.
- The system's dependence on software applications is kept to a minimum because the disk structure is open and is documented according to industry standards

- Objects stored in the permanent repository that have undergone changes through preservation actions are recorded as a version with auditing information
- Migration and media refreshment are easy to implement.

To ensure the long-term integrity of the information stored in the permanent repository, virus and fixity checks are run directly on the items stored in it.

Operational Repository

Besides the permanent repository, Rosetta provides an operational repository that enables searching, indexing, and quick access. The operational repository is where preservation actions take place. AIPs that are stored in the permanent repository and are identified as candidates for preservation action by the preservation planning module are subjected to the requisite preservation action in the operational repository. The output of the preservation activity is re-ingested into the permanent repository and stored as a new version of the original intellectual entity.

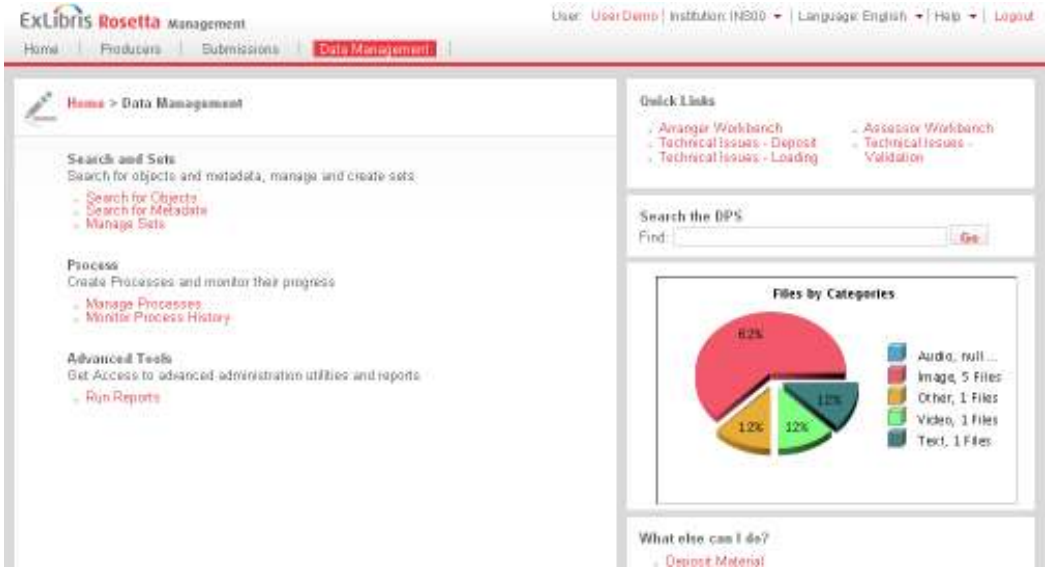

Figure 5. Data management

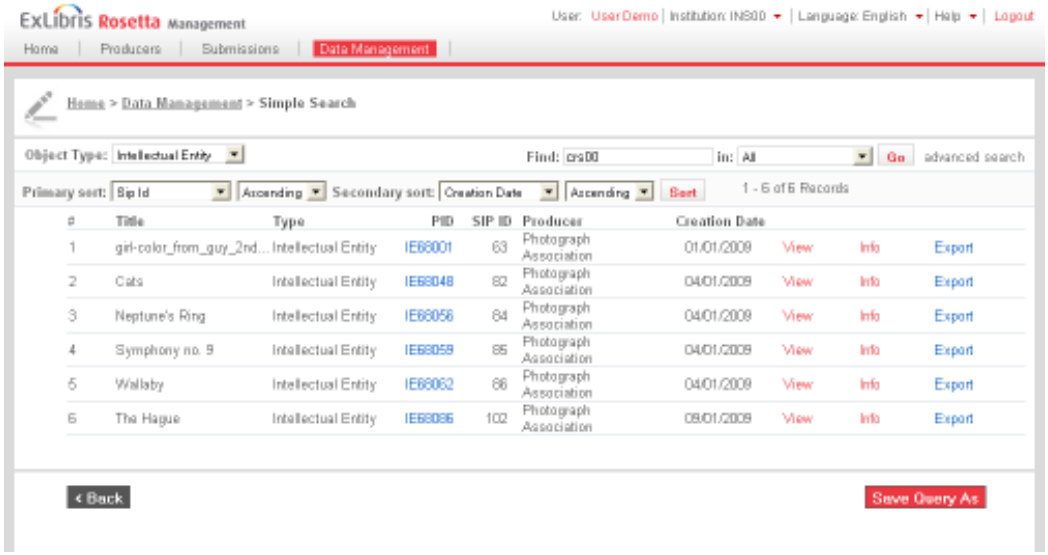

Figure 6. Results of a search for an object

Preservation Planning

The OAIS reference module specifies that a preservation system should include a preservation planning module. Rosetta includes a preservation planning module that comprises a format library[5], risk analysis, evaluation, and preservation actions. The module allows institutions to manage and carry out the full preservation process of identifying a risk, selecting the best among possible alternative solutions, and testing/activating the preferred preservation process.

1. Risk analysis: Based on format risks the system assesses the possibility of a format's becoming obsolete and provides a report on the risk to the repository manager. Risks can be added by the institution, but institutions can view risk analyses that derive from other institutions as well.

[5] A list of formats with relevant metadata on the different formats. The format library is based on PRONOM, a global format library maintained by the United Kingdom National Archives.

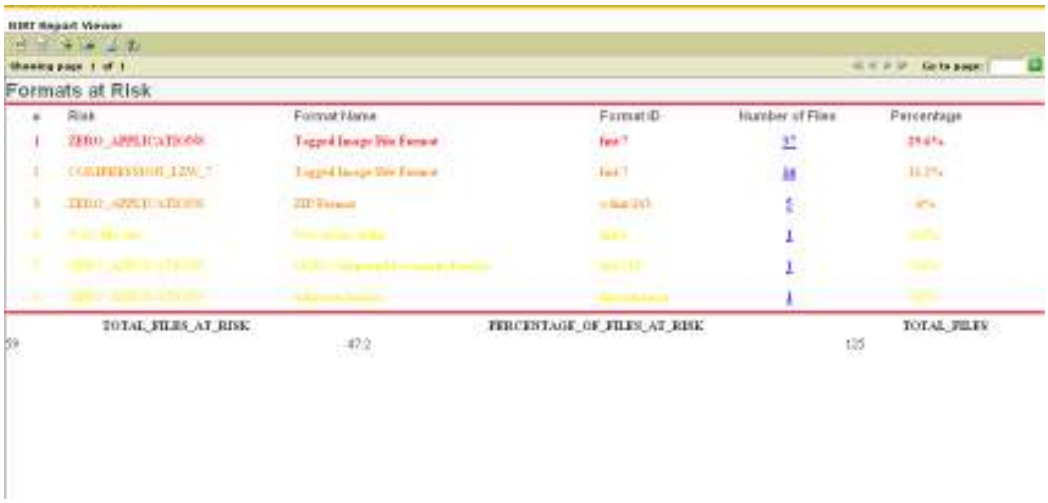

Figure 7. Risk report based on the files in the repository and the registered risks

2. Test bed: The institution can compare a number of preservation path options for materials that are identified as high risk. The system tests each of these options on a sample set of items and allows the institution to rate the success of the option based on parameters the institution has defined. Institutions can also view preservation plans from other institutions.

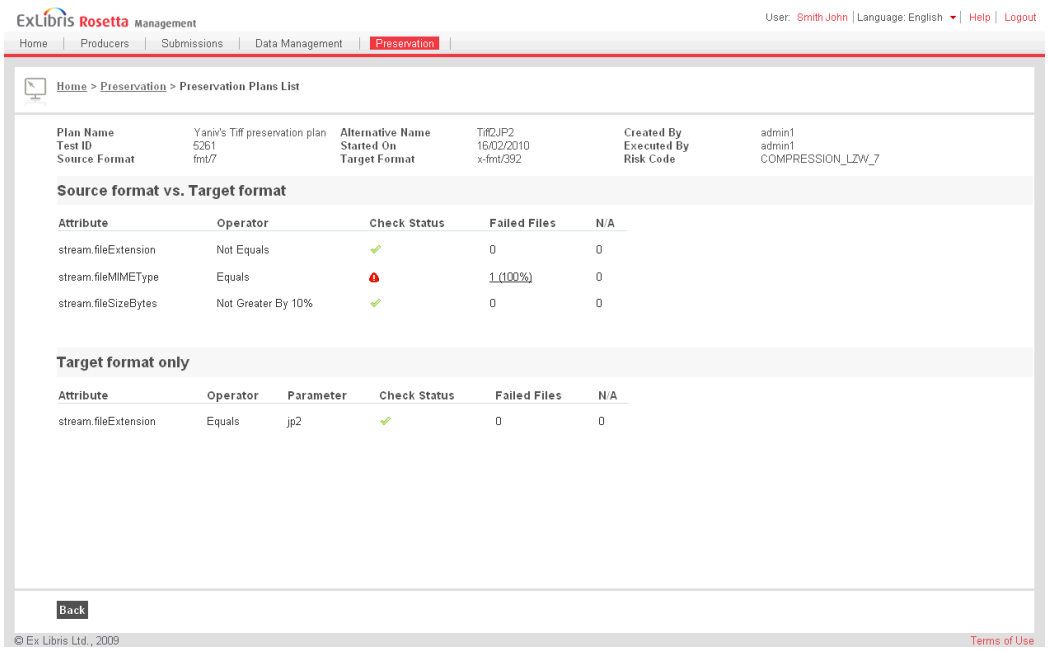

Figure 8. Review of the automatic comparison of the migration alternative.

3. Preservation action: After the institution has determined the best preservation approach for a given set of items, the system implements the preservation action on the entire set. Each preservation action that is

carried out is labeled with an ID and a description. Information regarding objects that are undergoing or have already undergone the preservation activity is stored in the preservation planning module. Thus, staff users can view the status of a preservation action and automatically run the same action on additional objects of the same format.

Administration

To support system management, the administration module includes the following cross-application functionalities, and more:

- System configuration: Users manage system configuration through a back-office Web-based, wizard-like interface.
- Reporting: Integrated into Ex Libris Rosetta is the open source, Eclipse-based Business Intelligence and Reporting Tools (BIRT) reporting system, which provides detailed lists and statistics on various aspects of the system and collections. Alternatively, the institution can choose to integrate an external reporting system, such as Crystal Reports.
- User management: Staff users, patrons, and producers can each have a set of customizable roles and privileges. The system supports integration with external identification, authentication, and management (IAM) systems via standard interfaces such as LDAP.
- Auditing: All activities performed on the digital objects maintained in the system are fully logged and traceable. Therefore, any changes that are made to an object after its ingestion into the system are recorded. Action information is stored in the system according to the effect that the action might have on the stored object.
- Monitoring: The system includes an interface that enables staff to monitor maintenance jobs and other processes that are running in the system.

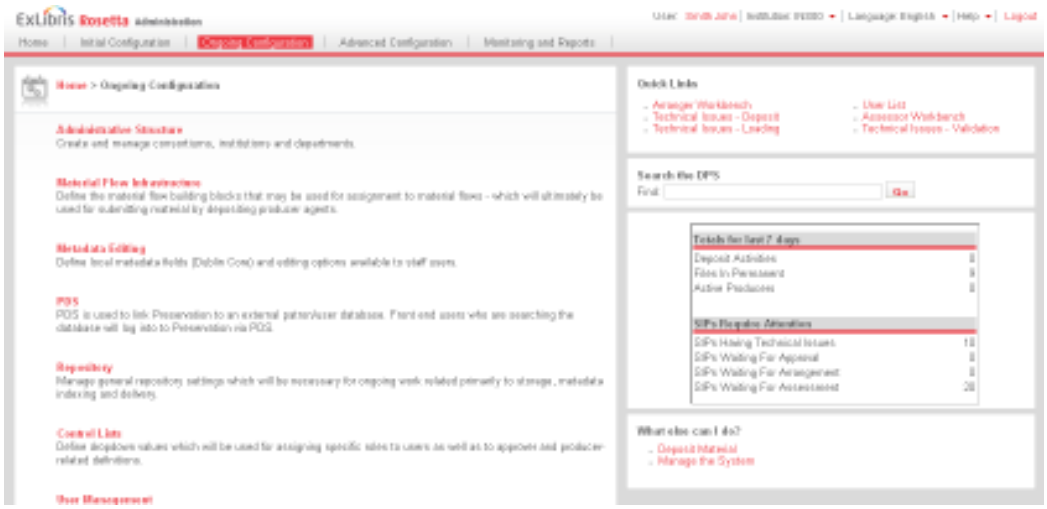

Figure 9. The Ongoing Configuration area in the administration module

Access

Dissemination information packages (DIPs) are created in the access module. DIPs are used for the dissemination of information stored in the permanent repository.

The access module consists of two components:

- The publishing component enables external systems (for example, search engines and resource discovery solutions such as Primo® from Ex Libris) to use standard communication protocols, such as OAI-PMH, Search/Retrieve via URL (SRU), and Search/Retrieve Web Service (SRW), to access data stored in the permanent repository.
- The delivery component delivers individual preserved intellectual entities via viewers that are supplied as part of the system (for example, a video streaming server) or as third-party applications. The system processes item delivery requests and checks the access rights defined for an item before delivering it to the user.

Figure 10. Example of a delivered item

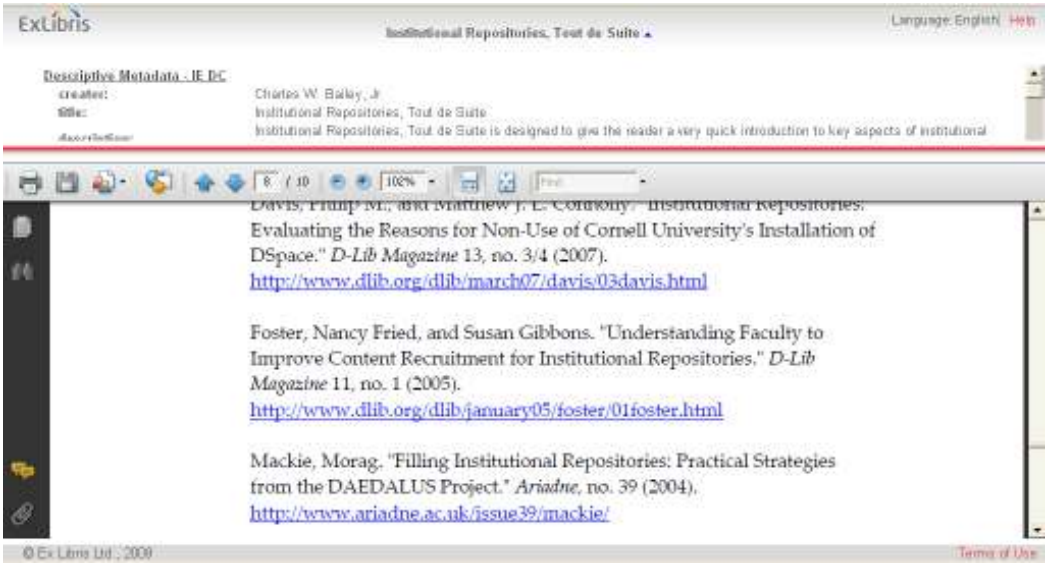

Figure 11. Example of a delivered item

Scalability

Ex Libris Rosetta provides a scalable infrastructure to address the ever-growing need to preserve and manage digital materials. This scalable solution can be implemented on a robust distributed architecture which allows the deposit module, working area, permanent repository, and database to be deployed on separate servers. Each module can be scaled up with additional of computers. In addition, Ex Libris Rosetta enables institutions to add dedicated servers, called *workers*, to perform specific tasks, such as virus and fixity checks. This flexibility allows an institution to start with a small hardware configuration and expand Rosetta to meet the needs of the institution's growing collection.

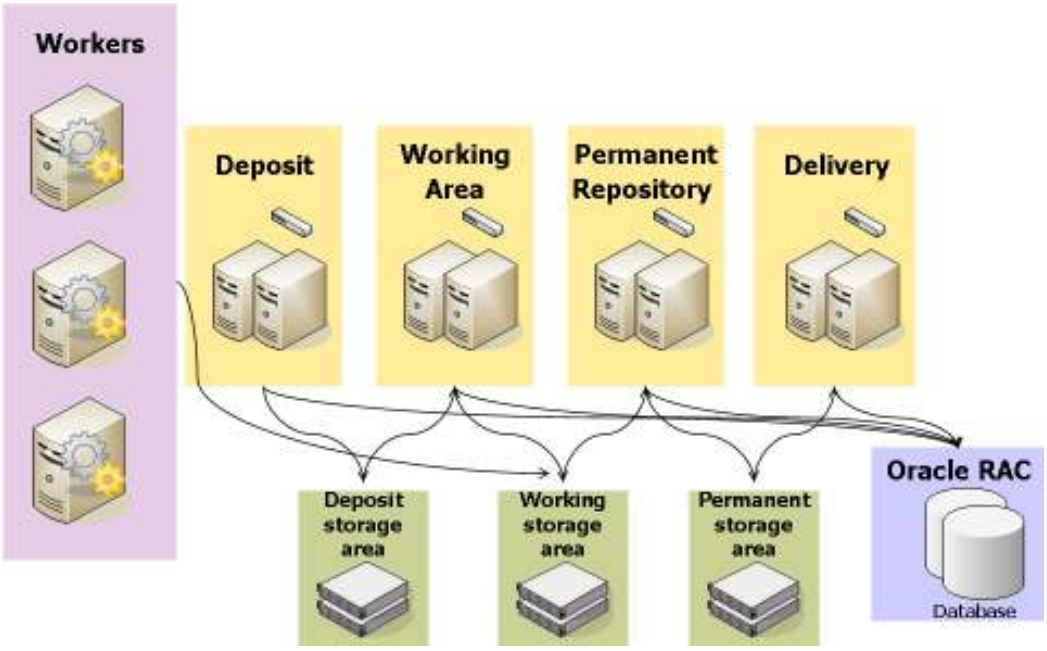

Figure 12. Scalability diagram

Beyond the benefit of scalability, the architecture provides institutions with a system that offers redundancy and has no single point of failure.

Openness

The Rosetta platform supports ingest, management, preservation, and delivery processes. Its open architecture combined with a rule-based workflow engine enables institutions to enhance the system with plug-in tools and applications in addition to those offered out-of-the-box. Such tools support the following tasks and activities:

- Submission: A set of APIs and a software development kit (SDK) enable institutions to create unique submission applications that support integration with existing institutional applications and can be seamlessly integrated with Rosetta.
- Characterization: External third-party tools such as JHOVE, DROID, and virus check applications are embedded in Rosetta to ensure that information in the system is viable and not damaged. New tools can be added as they emerge or evolve.
- Enrichment: Managed by a workflow engine, the rule-based ingest solution allows institutions to embed a wide variety of tools that add information to the AIPs stored in the system—for example, tools that convert file formats such as AVI to MP3. External metadata management systems (such as an ILS

catalog or an archival system) can be additional sources of metadata enrichment.

- Dissemination: The dissemination SDK is a set of tools that enables Rosetta to be integrated easily with local systems for digital rights management, e-commerce, and discovery.

Find Out More

For information about Ex Libris Rosetta, related events, and Webinar opportunities, go to the Ex Libris Web site or write to rosettachartergroup@exlibrisgroup.com. Product APIs are available for Ex Libris customers to download and use from the EL Commons Web site.