Reference Model, Requirements, and Reality: A Case Study on Implementing OAIS at OCLC

Pam Kircher and Dawn Lawson
OCLC, Online Computer Library Center
Dublin, Ohio

Abstract

This case study examines the development of the OCLC Digital Archive, a third-party service that provides (1) tools for the capture of individual online resources and offline collections; (2) a repository in which those resources and collections can be stored for preservation purposes; and (3) an administration module, which allows depositors to manage their archived resources after submission.

The OCLC Digital Archive complies with the *Reference Model for an Open Archival Information System (OAIS).* OAIS is a framework, implementations of which vary. The case study focuses on OCLC's development of requirements based on the OAIS and member input, highlighting factors that influenced our decisions.

Several categories of factors influenced the three-year development project. These factors include the nature of OCLC, the institution developing the archive; the local depositor community; and the global digital archiving community. Implementation decisions affected include object types and formats accepted into the archive, access methods, preservation metadata creation, types of tools developed, rights management capabilities, and preservation planning.

Introduction

In 2000 OCLC, an international, nonprofit membership cooperative, responded to member requests that it expand its preservation services by adding an emphasis on digital preservation. In support of that commitment, OCLC created a new division, Digital Collection and Preservation Services.[1] This new division was charged to develop and implement a digital archive.

The OCLC Digital Archive was implemented in phases over three years. The initial goals, in January 2001, were to build a general-purpose archive for digital objects based on the draft ISO standard *Reference Model for an Open Archival Information System (OAIS)*[2]; to identify workflows for capturing and managing digital objects; and to implement a metadata set for the archived objects. Our team of eight consisted of a product manager, a project manager, and six developers. Our budget was approximately $2 million and our target delivery date was March 2002. In that time we had to comprehend the OAIS model, determine the requirements of the projected depositors to our archive, choose hardware and application software, build middleware, determine the data storage structure, design interfaces, determine how to price the services, begin preservation planning, and develop a metadata set. Ultimately, we missed that first deadline by a few months—we went into production in May 2002—and we lacked some important features (disseminate, delete, and a formatted reports display). We had implemented a majority of the archive infrastructure, a harvester to facilitate the capture of web documents, and tools for the creation of digital preservation metadata and the management of the digital content before and after its ingest into the archive. We had also established plans for the second development phase, which included adding batch ingest mechanisms.

OCLC Digital Archive Features

The OCLC Digital Archive is OAIS-compliant, meaning that it includes the six functional areas specified in the reference model: ingest, archival storage, data management, administration, preservation planning, and access. However, as its name implies, OAIS is a framework, implementations of which vary. In details, our archive differs from other digital repositories in terms of the hardware and software it employs, the types of formats accepted, and particularly in the pre- and post-ingest tools available to its depositors. To provide context for the development decisions discussed later, a brief description of the archive follows.

It is a single archive with two ingest mechanisms: item-by-item from the web, a manual process; and in large quantities using an automated batch ingest process. The web archiving process uses OCLC Connexion, a browser-based metadata creation tool, as the front end of the process. To batch ingest, depositors send their digitized resources on CDs or tapes and OCLC ingests them into the archive automatically while also creating preservation metadata records for each object. An administration module provides functions for depositors to manage their resources after ingest—both those ingested manually with the web tools and those that were batch ingested.

Influences on Development Decisions

The decision to build an OAIS archive immediately gave rise to the need for many additional decisions—about hardware

and software, database architecture and indexing, and so on—the kinds of decisions required by many systems development projects. This paper focuses on a different set of decisions—those needed to implement the functions of a digital archive within a specific depositor community—and how those decisions were influenced by three factors. The factors are (1) the nature of the institution developing the archive, (2) the local depositor community, and (3) the activities and emerging standards of the global digital archiving community. These three factors continue to shape the ongoing development of the OCLC Digital Archive.

Institution Developing the Archive

That OCLC is an international membership cooperative influenced what we built, when we built it, and how we built it. The goal was to develop an archive that could be used by OCLC's entire membership, which encompasses a variety of libraries and other cultural heritage institutions who hold diverse digital collections. From the beginning of the project, we were assisted by a small group of OCLC members designated as pilot depositors to the system. They constituted our local depositor community and are discussed below.

When it decided to build the archive, OCLC committed sufficient staff and financial resources to enable its development to be sustained over time. This meant that each development decision had to be feasible both in the present and as far into the future as we could imagine. This requirement was daunting and at times verged on the paralyzing. The difficulty wasn't just making the decision once we had the information to weigh and compare alternatives. The difficulty was finding information to weigh and compare. There were no preservation metadata standards, no studies of the use of archived objects over time, and no adequate cost analysis of digital preservation, or even a consensus on what preservation actions to take when and for which formats. The desire to make the right decision for the long-term was sometimes at odds with making a decision that would allow the project to stay on schedule. We sometimes had to remind ourselves that to take a step was the only choice we had.

We had the financial support of OCLC, embodied by a new division, staff, and a budget. Yet we needed to build understanding and commitment throughout the organization. We undertook an educational campaign, through presentations, to gain institutional support and acceptance. This was critical, as over a dozen work units are involved in enabling the archive to fulfill all the functions of an OAIS—from legal and corporate security, to order processing, to systems planning and operations.

Local Depositor Community

During development, the local depositor community consisted of the OCLC members likely to be the initial and early depositors to the archive, including four US state libraries, the US Government Printing Office (GPO), and the University of Edinburgh. These members guided our development requirements and assisted in usability assessments. They participated in meetings, commented on prototypes and workflows, and provided input into the metadata element set.

From the outset, the pilot institutions and OCLC agreed that the OAIS was the model to follow for the archival functions. However, we soon realized that our project scope went beyond the OAIS requirements. In particular, pre-ingest and post-ingest management of objects is of great importance to our depositors. They needed us to automate and facilitate the processes of selecting, harvesting, and archiving. In addition, we talked extensively with them about workflow, staffing, types of materials, reports, and other details. We visited one institution and interviewed others by phone to gather requirements. However, many of the digital archiving processes under discussion were new to the institutions and workgroups; therefore they—and we—often could only guess at what would be needed.

The make-up of our pilot group influenced the object types and formats accepted by the archive as well as the tools needed pre- and post-ingest to manage their collections. All of the institutions wanted to capture web objects; most had legal mandates for preservation that they were just beginning to apply to born-digital material. The basic requirements for selection, acquisition, cataloging, and preservation were common to all, but at the same time the institutions were distinct enough that the processes they used to accomplish these tasks differed, which meant that the tools we developed had to accommodate a variety of workflows.

After a few months of discussion with the pilots, we had collected many requirements and had the rough outline of a system in place. It had also become clear that we would not be able to meet all the suggested requirements by our scheduled production date. Requirements removed from the original project, such as tools that would alert depositors to changes on a selected web site or to check for duplicates before ingesting an object, are still desirable today, and development continues. Each month brings new enhancements to the archive system.

Global Digital Archiving Community

Throughout our development process, the global digital archiving community provided a source of inspiration, practical experience, and evolving standards from which to draw. The nature of the digital preservation problem is such that only a community of repositories can resolve it, and that community is indeed where the standards for digital repositories, metadata, and interchange are evolving. There will be multiple repositories, and interchange of objects among them will be a necessity. Hence, any digital archive builder has a responsibility to understand and apply the standards, and to the extent feasible, participate in their creation, development, and maintenance.

In addition to our reliance on the OAIS model, we utilized the project information available from CEDARS, NEDLIB, and PANDORA. We read about research in publications by CLIR, the Digital Library Federation, and in *RLG DigiNews* and *D-Lib Magazine*. We also followed

conference proceedings and ongoing research projects at academic institutions such as Harvard, University of California at Berkeley, and MIT.

The activities of the global digital archiving community continue to exert a strong influence on the development of our archive. The emerging Metadata Encoding and Transmission Standard (METS) provides an XML schema for encoding metadata with a digital object. In 2002 the draft NISO Technical Metadata for Digital Still Images standard was released and a METS extension, MIX, was created by the Library of Congress to support it. New York University developed textMD, a METS extension for encoding technical metadata for text.

As follow-ons to the OAIS, OCLC and RLG convened two international working groups. One working group resulted in the publication of a *Metadata Framework to Support the Preservation of Digital Objects,* which describes the metadata necessary to carry out, describe, and evaluate digital preservation processes.[3] The second working group produced *Trusted Digital Repositories: Attributes and Responsibilities.*[4]

In 2003 two new international working groups formed. OCLC and RLG sponsor PREMIS (Preservation Metadata: Implementation Strategies), the goal of which is to create a practical, implementable preservation metadata set. RLG and NARA convened the Task Force on Digital Repository Certification to produce certification requirements for establishing and selecting reliable digital information repositories.

Decisions Influenced

Among the decisions influenced by the three factors discussed above are the formats and object types accepted into the archive, access methods, preservation metadata created by depositors and our system, types of tools developed, rights management capabilities, and preservation policies. The decisions discussed below are only a fraction of the myriad decisions made overall, and only a few of those influenced by the factors. Further, one decision affected the next decision, as described in the section on formats and object types, which in turn influenced decisions on preservation metadata, rights, and access methods.

Object Types and Formats

Early in the development cycle the local community reached consensus on the object formats to start with: text and still image. They knew from their previous work with web documents that these predominated—for now. Also, we believed them to be fairly stable formats.

We currently accept the following: PDF, HTML and associated java script and cascading style sheets, TXT, TIFF, GIF, JPEG, and BMP. This list of formats was based both on our assessment of the likelihood that we would eventually be able to offer full preservation of content in these formats and on our depositors' immediate needs. OCLC plans to expand the list of accepted formats based on depositor request and preservation potential.

These decisions about object types and formats clearly affected the tools we built, such as the harvester; the first method of access, a browser; and the amount of rights information preserved with each object.

Access

OCLC and its local depositor community decided that the WorldCat (OCLC Online Union Catalog) record or other bibliographic record would be the end-user's discovery and access mechanism. That bibliographic record might be in a library's catalog, or in a reference service like OCLC FirstSearch. This approach makes sense given the amount of use WorldCat and other catalogs receive each day, the sophistication of searching capabilities in WorldCat and local library systems, and that, for our local community, the archiving workflow is associated with a cataloging workflow.

This decision led to another. We chose to build on existing OCLC authorization structures. In our administration module, we included access controls based on the depositor's existing OCLC cataloging and FirstSearch authorizations. We also included an IP recognition option based on the capability already available in FirstSearch. Although these are by no means ultimate solutions, they work well for now because most of the objects deposited (government documents) are in the public domain. However, using their existing OCLC authorizations or IP recognition, depositors can choose to permit local access only, meaning that they can view it, but no one else. For example, TIF images may be stored in the archive and accessible only by the depositor while derivatives are made available for public access on the web. At OCLC, we differentiate between disseminating (delivering both object and metadata) and viewing (delivering the object only). The depositors of the objects can disseminate both object and metadata whenever they like. Because the first objects in the archive were web documents, we rely on current browsers to be able to render the objects delivered without additional metadata, structural or administrative.

The decisions described above illustrate how the nature of the developing institution and the depositor community influenced our development. Because OCLC is developing the archive, we made use of existing infrastructure such as WorldCat and FirstSearch. Our local community already used these tools and welcomed the expansion of their functions to digital archive processes. The decision to use the WorldCat record for discovery also influenced preservation metadata decisions, as described below.

Preservation Metadata

Although international collaboration toward that goal is ongoing, as mentioned above, there is not yet a standard preservation metadata set. At the time that our archive went into production, the OCLC/RLG Working Group on Preservation Metadata had not released its report. Instead, the preservation metadata set implemented in our archive was developed by OCLC staff and our local depositor community. We drew heavily on the preservation metadata

work of CEDARS, NEDLIB, and National Library of Australia. Also, we were able to review early drafts of the OCLC/RLG framework document.

During the initial development phase we frequently reminded ourselves that we could not have all the answers right then and that the metadata set would evolve. And it has. At present we have about forty elements.[7] Some elements have become obsolete and some new ones have been added in the past year. We know the preservation metadata will change further as a consequence of changes to the archive's functionality, such as the addition of new ways to ingest objects; the expansion of the number of object types accepted into archive; and the international community's establishment of a standard for preservation metadata.

As mentioned, OCLC and the pilot participants began by looking at the elements proposed by other projects from the global community. We developed questions to help us to consider each element, including: was the element needed for the types of objects we were archiving; did we understand the proposed element's purpose and scope; when would the information contained in the element be available for capture, extraction, or creation; and for what functions would the depositor or the system need the metadata? In answering these questions, we found that metadata decisions are also influenced by characteristics of the local depositor community, by the processes that create and use the metadata, and by characteristics of the objects.

Characteristics of the Local Community

The web objects that our local community is responsible for preserving were not created by them. They may not know or be able to obtain some preservation metadata elements such as object origin information, object composition, or the relationships among the files. As a result, our system does not require the depositor to supply much metadata at ingest: simply the title of the object, the language of the object and the metadata, and a bibliographic record number. The system then creates or extracts other required metadata, such as the object composition and the URL map, which records the relationships of files to each other in a web object.

As mentioned, we decided that the bibliographic record, not the archive record, would be used by patrons for resource discovery. The archive metadata is utilized by the depositors of the objects and by the system for management and preservation. Accordingly, we tried to limit the amount of descriptive metadata in the archive to the amount needed by a human to identify the object and distinguish it from other similar objects. For example, we do not include subject headings in our preservation metadata.

At the outset, our local community did not want to see all the technical metadata our system extracts. They were uncertain as to how they would use it in their processes and procedures, although they understood its importance for preservation purposes. However, this decision has now been reconsidered by the local community, and as a result, we will make additional technical metadata available for display in the near future.

Processes to Create and Use the Metadata

Our local community wants to integrate workflows: to select, capture, catalog, and archive in a streamlined fashion. They need the process to be straightforward, so that staff of all levels of technical ability and experience can ingest objects into the archive and still obtain the necessary preservation metadata. Building our metadata creation tool on the OCLC Connexion interface (discussed below) allowed us to integrate cataloging with archiving, including mapping certain fields used in both the bibliographic and preservation metadata records from one to the other.

Not only did we consider the ease with which a depositor might create the metadata, but whether tools existed to help them and even whether the element was already widely understood or whether it was new and unfamiliar. An example of the latter is "Significant Properties." We believe significant properties will be important in the future, but at the time we were implementing our metadata set there wasn't any consensus on how this concept applies within a given community and to particular object types.

Characteristics of the Objects

Clearly, technical metadata requirements are determined by the object and are fundamental to preservation. The requirements for specific types, such as text, image, and audio, are being developed by experts at institutions around the world. We intend to rely on that work for determining much of the technical metadata needed by our archive.

Other metadata that could be provided by the depositors of the objects is influenced by the object type. As mentioned, our local community focused on archiving born-digital web documents that they did not create; therefore, "Object origin" is one of the fields that the depositors did not find important to their work. Further, they were harvesting freely available web documents, which are not encrypted or password protected; therefore, we did not implement "Access inhibitors." Finally, these were public-domain government documents, so the rights information they felt necessary was minimal (see below).

Tools

We created five major tools to assist with the pre-ingest and post-ingest administrative functions of the archive. The tools derived from the workflows established by the local community, which in turn derived from the object types and formats. Development of the tools was further affected by the metadata decisions. The five tools are the metadata creation tool, the harvester, the administration module, the submission builder, and the dissemination interface. Development of each was influenced by issues of complexity versus ease of use, performance, and usability, among others.

As mentioned, we built the metadata creation tool into Connexion, OCLC's browser-based metadata creation tool. This was a good decision in terms of ease of use and workflow flexibility and compatibility; it provides a level of commonality among cataloging and archiving tasks, and offers a familiar interface in which to do new tasks. One of

our depositors, the Connecticut State Library, used this tool to create metadata for the more than 3,000 objects that they ingested in one year. They used Connexion to export each record from WorldCat to their local OPAC; as a result, each of those objects is accessible from a link in the local bibliographic record.

However, developing for the Connexion platform was challenging at times. The archive is one of several services supported by Connexion; therefore we could not always change the platform to accommodate our processes. In fact, in some cases the Connexion infrastructure determined how our process runs. For example, ingest is a synchronous, rather than asynchronous, process due to the Connexion relationship. In addition, the relationship between the Connexion cataloging and the archive interface also means that as changes are made to Connexion, we need to assess the impacts of these enhancements on the archive, decide whether to implement them—with any necessary adjustments—and then coordinate the testing and implementation.

The harvester bridges Connexion and the archive, adding additional complexity to the development process. The harvester communicates with both systems, and must return status messages to Connexion without delaying Connexion processes. The harvester provides substantial information about the web object's composition prior to the actual harvest, allowing the depositor to determine the boundaries of the object to be harvested. The amount of information returned, the display of the information, as well as explanations of the harvester's actions (such as inclusion and exclusion of files) has been the subject of continual examination and enhancements.

We built an administration module to allow the depositor to set access permissions on an object, to manage collections via content groups, to create and apply rights statements, to view reports, and to view objects for quality assurance. We recently released a complete redesign of this module that addresses scalability issues such as retrieval speed.

We also built tools to create METS-encoded packages of metadata and objects. These are the submission builder and the dissemination tool. Technical metadata about each file is included in the DIP (Dissemination Information Package) in either the MIX or textMD formats. The submission builder is a stand-alone PC-based application; the dissemination function is part of the Connexion interface.

Rights Management Capabilities

Our decisions about rights management capabilities were driven by the local community and the types of objects they most needed to archive. As mentioned, the local community needed to preserve government publications to allow public access over time. Copyright is not an issue with these publications. Further, we rely on the depositor to ingest objects for which they have legal responsibility and explicitly state in our submission agreement that the depositor has that responsibility to comply with copyright. Therefore we created a mechanism to associate a simple rights statement with each object. Additional mechanisms allow the depositors to restrict access if needed. It is clear that our archive must eventually accommodate more complex rights needs, but because our initial development made provisions for basic rights statements and access controls, we are able to wait while the global community makes progress on digital rights metadata via projects such as XrML, extensible rights markup language; ODRL, open digital rights language initiative; METSRights, the schema for rights declarations; and PREMIS.

Preservation Planning

The first phase of development implemented bit preservation services for all objects in the archive. Since then we've been developing an approach to full preservation. The OAIS describes a preservation planning entity, which monitors the digital archive environment and ensures that the content objects in the digital archive remain accessible over the long term. The preservation planning entity evaluates the contents of the archive, suggests preservation actions, recommends standards and policies, and monitors the technology environment and the designated community for any changes in its service requirements.

OCLC will engage in risk assessment to detect the timing and likelihood of changes in the overall technology environment and in individual software formats that will affect accessibility and long-term preservation. Each format accepted into the archive will be risk-assessed in these areas: format, required software, required hardware, and associated organizations. Each format will have a detailed preservation plan, which will describe the preservation approach and define immediate, intermediate, and long-term preservation actions. Preservation plans will be reviewed periodically and updated.

A major consideration in preservation planning is taking actions that support the local depositor community's long-term needs for use of the digital object. OCLC plans to incorporate local and global community input into its preservation planning process. One way is through a monitoring process known as "technology watch," which is designed to forecast and assess future changes in technologies and emerging trends that have potential impacts on long-term access and preservation. The findings of technology watch activities include detailed information regarding digital format specifications and software and hardware components. It is expected that technology watch will evolve into a cooperative effort engaging the entire international digital preservation community.

OCLC will make its preservation policy available to the public on its web site in the near future. The document—and the policies explained in it—will be dynamic in nature and therefore will be updated frequently. We will encourage comments on and questions about our preservation processes.

Conclusion

The OAIS is a useful framework from which to begin building a digital archive. However, every repository

implementation inevitably will be influenced by characteristics of the institution building the archive, the local community, and the activities of the global digital archiving community. These three factors influence decisions about formats and object types, access, preservation metadata, tools, rights management capabilities, and preservation planning. At OCLC, the decisions—to use the bibliographic record for discovery; to build on existing infrastructures, such as Connexion and WorldCat; and to limit the accepted object types and formats—provided boundaries for the initial phases of digital archive development. A digital repository is a complex system; communicating with the local community and utilizing standards assist in completing the project successfully.

References

1. OCLC Digital Collection and Preservation Services. http://www.oclc.org/services/preservation/default.htm.
2. Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS).* Washington, D.C.: CCSDS Secretariat, 2002. http://www classic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf.
3. OCLC/RLG Working Group on Preservation Metadata. *A Metadata Framework to Support the Preservation of Digital Objects.* 2002. http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.
4. RLG/OCLC Working Group on Digital Archive Attributes. *Trusted Digital Repositories: Attributes and Responsibilities.* 2002. http://www.rlg.org/longterm/repositories.pdf.
5. PREMIS (Preservation Metadata: Implementation Strategy). http://www.oclc.org/research/projects/pmwg/default.htm.
6. Task Force on Digital Repository Certification. http://www.rlg.org/longterm/certification.html.
7. OCLC preservation metadata set. http://www.oclc.org/support/documentation/digitalarchive/da_metadata_elements/default.htm.

Biographies

Pam Kircher is product manager for the OCLC Digital Archive. In her 15 years with OCLC she has held positions in OCLC's cataloging and reference services units. Her MLS is from Kent State University. She has given presentations about the OCLC Digital Archive in many venues, including meetings of the American Library Association, Coalition for Networked Information, and the Society of American Archivists. Pam represents OCLC on PREMIS, the OCLC/RLG working group on preservation metadata.

Dawn Lawson was product manager for the OCLC Digital Archive in 2002-2003, prior to which she was product manager for electronic versions of the Dewey Decimal Classification® system at OCLC. After co-authoring this paper, she joined the New York University Division of Libraries as East Asian Studies librarian. Dawn holds a MA in Japanese literature from Harvard University and an MLS from the Palmer School of Long Island University.