Enduring Access to Digitized Books: Organizational and Technical Framework

Oya Y. Rieger
Cornell University Library
Ithaca, NY, USA
oyr1@cornell.edu

Bill Kehoe
Cornell University Library
Ithaca, NY, USA
wrk1@cornell.edu

Abstract

The digitization of millions of books under corporate and non-profit programs is dramatically expanding our ability to search, discover, and retrieve published materials. Accompanying this progress are cultural heritage institutions' concerns about the long-term management challenges associated with providing enduring access to a large corpus of digitized materials, especially within the confinements of copyright laws. The goal of this presentation is to describe Cornell University Library's program to illustrate a range of organizational and technical issues involved in planning and implementing a preservation infrastructure for digitized books.

Large-scale digitization of published materials has brought millions of books hidden in library stacks to the public eye, making them easy to identify and locate. During 2006-2007, when Cornell University Library (CUL) signed contracts with Microsoft and Google to embark on two large-scale digitization initiatives, the Library staff was equally excited and anxious about the new roles and responsibilities required to successfully manage such a program.

The Library has been involved in various digitization initiatives since the early 1990s; however, given limited funding and the available digitization technologies, CUL had managed to digitize only close to 12,000 books by 2006. At this rate, it would have taken us hundreds of years to convert our entire collection of 7 million items. Whereas the Microsoft collaboration, which lasted for 18 months, resulted in the digitization of close to 100,000 public domain books.

The Google digitization collaboration, which is still in the initial planning stages, involves digitizing approximately 120,000 books per year for five years, covering both public domain and in-copyright materials. In addition, although at a significantly lower pace, there is an in-house digitization operation that grew out of the Microsoft collaboration to systematically digitize special and rare materials from the Library's collection. The goal of this article is to describe the preservation infrastructure under development that will ensure the effective management of these digital assets.

Preservation Framework

The Cornell University Library drafted its first digital preservation policy framework in 2004, formalizing the library administration's ongoing commitment to the long-term preservation of its diverse digital assets. Although a strong mandate was articulated and the policy included a range of operating principles, roles, and responsibilities, the policy did not move into an implementation stage until the launching of the large-scale digitization initiatives. The prospect of assuming the responsibility of a large body of digital content prompted the library staff to take quick steps to develop a preservation program.

The three legs of the Cornell digital preservation program include *organizational framework*, *technological infrastructure*, and *resource requirements*. Utilizing this three-tiered approach, the following sections describe the decision-making and implementation processes for CUL's preservation program for digitized books. The original three-tiered approach has been expanded to incorporate *access mandate*, which has a critical value for current and future scholarship.

Organizational Framework and Policy

Throughout the last 15 years, we have learned from first-hand experience that technologies alone cannot solve preservation problems. Institutional culture, policies, strategies, staff skills, and funding models are equally important. Organizational infrastructure includes policies, procedures, practices, people – the elements that any programmatic area needs to thrive, but specialized to address digital preservation requirements.

Digital preservation requires a sequence of decisions and actions that begin early in the life cycle of an information object. Standard policies and operating principles for digital content creation are the foundation of a successful preservation program. The critical components include:

- Technical specifications for content creation to specify image-quality parameters for archival and derivative files;
- Requisite preservation metadata with descriptive, administrative, structural, and technical information to enhance access, enable content management, and facilitate discovery and interoperability;
- Quality control and assurance protocols for digital images and associated data.

Although the Library had established digitization and metadata standards prior to the initiation of the large-scale conversion project, we had to reassess our requirements within the scope of our collaborations with Microsoft and Google. Due to the collaborative nature of the initiatives, the companies' digitization protocols and target outcomes set the parameters for digital content creation process.

As the Library was negotiating the contracts with Microsoft and Google, the University Librarian appointed a team called Large-Scale Digitization Steering Committee to oversee various phases of the initiatives with a holistic approach, from selection and preparation of materials to ingest and archiving of digital books. In addition, the Committee was charged with the critical process of identifying staff skills and patterns (and associated costs) required to implement digitization and preservation strategies. One of the Committee's first challenges was to define a new set of requirements that could be supported by the technical provisions of the corporate partners – to compromise between what was available with what was desirable. Some of these technical decisions are illustrated in the following section.

An example from the Committee's current agenda involves exploring our legal rights to preserve in-copyright content. Although the Library's Microsoft project focused on public-domain materials, the collaboration with Google includes 500,000 books representing both in- and out-of-copyright materials. We have a myriad of question to address. For example, is it legally permissible for a library to rescan originals that are not in the public domain to replace unusable or corrupted digital objects? What are the copyright implications of migrating digital versions of materials in copyright from the TIFF to JPEG2000 file format? Section 108 of the U.S. Copyright Law articulates the rights to and limitations on reproduction by libraries and archives; however, the right to take action to preserve digitized content that is copyright protected is still under study by the Section 108 Study Group convened by the Library of Congress.

Technological Infrastructure

E-science data initiatives have introduced libraries to the challenges associated with large-scale database storage and retrieval. Nonetheless, many participating libraries still have limited experience in data management at the scale of these initiatives, even though the technology that makes preservation possible has the same basic components as the technology of digital collections. The following sections highlight some of the important components of our technological infrastructure, especially from decision-making perspectives.

JPEG2000 as an Archival File Format

The page image files in our digital archive constitute 97 percent of the space required to store the digital books. The format used for storing the images has become important not only from the perspective of best practice for digital preservation, but also from the economic view of sustainability over the long term. Fortunately, best practice and fiscal prudence meet in the JPEG2000 format. Others have reported on the archival benefits of the format—for example, its capacity to embed metadata and yield scaled derivatives easily. Lastly, its ability to be compressed without significant visual degradation translates into significantly lower storage costs.

Physical Storage

For most of its servers, the Library contracts with Cornell's central information technologies group for maintenance and storage. That arrangement proved most cost-effective when we investigated the options for large-scale storage. At the beginning of our search, we expected to store JPEG page images and assumed a need for about 100 terabytes. Our decision to convert the JPEGs to the JPEG 2000 format reduced our storage need by more than 60 percent, and a 40-terabyte array of 1-terabyte SATA drives from Digi-Data Corporation satisfied our requirements for a unit of storage. One unit was sufficient for the first year of production (although we expect to make additional unit purchases in the coming years). The disks are being managed on a three-year lifecycle as a write-once array, in order to minimize maintenance. Deletions are discouraged—a maintenance policy that is easily met by our preservation policy, which demands that nothing be deleted and that any updated objects are added as new versions of earlier objects.

Redundancy Arrangements
Backing up terabytes of data to tape, even static terabytes that aren't expected to change, is a slow, cumbersome process. Restoring a large-scale system from tape would also be very slow. The Library has chosen to assure redundancy by keeping copies of the archived objects on remote storage arrays. Partners with access to Internet2 can speed copies to us if necessary. To mitigate the risk of losing our metadata, however, the XML containers are being backed up to tape locally.

The Choice of an Archival Storage System
After having decided that we would not build a data management and archival storage application ourselves, we examined the characteristics of aDORe and Fedora. We set up test implementations of each and experimented informally with ingest and access. Both systems showed themselves to be capable of managing complex objects well. At the time we investigated the systems, Fedora was the more flexibly access-oriented of the two, while aDORe had the more stable indexing mechanism for an object's component files. Even though Fedora's large user community and its flexible object model were very attractive, aDORe's storage model—its use of the Internet Archive's ARC-file format and cross-indexed XML metadata containers—promised to use our storage array more efficiently. With our primary focus on the archiving our digitized books rather than providing public access to them, we chose to base our system on aDORe. Nevertheless, we appreciate Fedora's capabilities and plan to use it as the middleware framework for a user-oriented access system as well as reassessing our decision to use aDORe.

Illustration 1: High-level view of the aDORe Archive system

(from http://african.lanl.gov/aDORe/projects/adoreArchive/; used by permission of Los Alamos Nation Laboratory Research Library

Archival Storage Architecture
The Los Alamos National Library's aDORe archive is a self-contained archival storage system based on the OAIS Reference model. The core is a dual-format storage mechanism: Metadata about complex objects is aggregated in a format called XMLTape; the datastreams that constitute the objects' files are stored in the ARC file format originated at the Internet Archive. The OpenURL's pointing to the datastreams are indexed for ease of retrieval. References to the datastreams are embedded in the XMLTapes. An index of identifiers and timestamps enables OAI-PMH access to the data through the XMLTapes.

Objects to be ingested must first be described in an XML format; Cornell uses a METS container. An external database is used to provide mapping between Descriptive Metada and aDORe OpenURLs for administrative and user access.

Metadata Requirements

Preservation metadata incorporates a number of emphasizes recording digital provenance (the history of an object). Documenting the attributes of digitized materials in a consistent way makes it possible to identify the provenance of an item as well as the terms and conditions that govern its distribution and use.

The role of technical metadata (or lack thereof) in facilitating preservation activities is not yet well documented. Although incorporated in preservation metadata, technical metadata merits special mention because of its role in supporting preservation actions. Published in 2006, ANSI/NISO Z39.87 Technical Metadata for Still Images lays out a set of metadata elements to facilitate interoperability among systems, services, and software as well as to support continuing access to and long-term management of digital image collections. It includes information about basic image parameters, image quality, and the history of change in document processes applied to image data over the life cycle. The strength and weakness of Z39.87 is its comprehensive nature. Although in many ways an ideal framework, it is also complex and expensive to implement, especially at image level. While most of the technical metadata can be extracted from the image file itself, some data elements relating to image production are not inherent in the file and need to be added to the preservation metadata record.

It is difficult to consider an image to be of high quality unless there is requisite metadata to support identification, access, discovery, and management of digital objects. Descriptive metadata ensures that users can easily locate, retrieve, and authenticate collections. CUL relies on bibliographic records extracted from local Online Public Access Catalogs (OPAC) for descriptive metadata. Compared with early digitization initiatives, minimal structural metadata are captured. We are committed to use of a persistent IDs to ensure that globally unique IDs are assigned to digitized books; however, we have not yet developed an access system to address this requirement. We do not capture detailed structural metadata, which facilitates navigation and presentation by providing information about the internal structure of resources, including page, section, chapter numbering, indexes, and table of contents.

Resource Requirements: Understanding Financial Implications

Some digitization costs such as materials shipping, scanning, processing, OCR creation, and indexing are covered by Microsoft and Google. However, staff members at the Library are supporting these initiatives by spending significant amounts of time negotiating, categories, including descriptive, administrative and structural. PREMIS metadata planning, overseeing, selecting, creating pick lists, extracting bibliographic data, pulling and re-shelving books, and receiving and managing digital content. This is an exhausting and disruptive workflow, and its associated local expenses are significant.

During Fiscal Year 2008, Cornell University Library invested close to seven full-time equivalent staff (distributed among a total of 25 staff members) in managing LSDI-related tasks for digitizing 10,000 books a month. It is difficult to calculate a fixed cost because of individual factors that affect selection and material-preparation workflows and the varied physical environments at participating institutions. Different staffing configurations are also required for ramp-up versus ongoing processes. Often neglected or underestimated in cost analysis are the accumulated investments that libraries have made in selecting, purchasing, housing, and preserving their collections.

Although our initial preservation strategy is comprehensive and treats all the digitized books equally, one of the questions we need to explore is whether we should commit to preserve all the digital materials equally, or implement a selection process to identify what *needs* to be preserved, or assign levels of archival efforts that match use level. According to a widely cited statistic, 20 percent of a collection accounts for 80 percent of its circulation. An analysis of circulation records for materials chosen for Cornell University Library's Microsoft initiative showed that 78 percent to 90 percent of those items had not circulated in the last 17 years. In Cornell's case, the circulation frequency may be lower than average because of the age of the materials sampled: all were published before 1923.

Because selection for preservation can be time-consuming and expensive, the trend will likely be to preserve everything for "just-in-case" use. The long-tail principle also may prove that every book finds its own user when it is digitized and discoverable on the Web.

Access Mandate

The 800-pound gorilla in the Library's preservation agenda is the future of Web access to digitized books. Several staff members expressed concerns that digital content may no longer be available in the future through present-day search engine portals, which evolve rapidly in terms of both content and retrieval technologies.

The May 2008 announcement about the closure of the Microsoft Live Search Program proved that the apprehension was not unwarranted. The Microsoft Live Book search website was closed down as soon as the

announcement. Because the Library was relying on using the Persistent IDs provided by Microsoft to connect users from its online catalog to digital books, the unexpected development caused a reroute to square one in means of exploring access options.

Currently, the Library has plans in place to implement bit preservation. However, providing enduring access by enabling online discovery and retrieval of materials (within limitations of copyright laws) for future generations is an enormous challenge—one that may not be met unless faced collectively by research libraries. Efforts at the individual library level will not adequately address the enduring-access challenge unless there is a plan for providing aggregated or federated access to digital content.

From scholarship perspective, the scale of the digitization undertakings is exhilarating and introduces the possibility of novel ways of finding and analyzing content that have been historically presented in print formats. Today's users prefer searching and retrieving information in integrated search frameworks and use digitized books only if they are conveniently accessed at their preferred search environments and support their searching and reading preferences. Therefore, hosting public domain digitized books solely through individual library portals is likely to be insufficient. Having more than one search engine host the same content is likely to increase the survival of digital materials.

Although today's users typically prefer to search for resources online, recent surveys and anecdotal evidence suggest that many users continue to favor a print version for reading and studying—especially for longer materials such as books. This is especially true for humanists as their scholarship heavily relies on close reading and interpretation of texts. CUL has been using the print-on-demand service provided by Amazon/BookSurge to make digital content created through institutional efforts available for online ordering. Thus far the initiative has been limited to the books digitized through past digitization initiatives. The Library is in the process of assessing the PoD options for public domain materials digitized through Microsoft collaboration.

Concluding Remarks

Large-scale digitization initiatives have been unexpected and disruptive—at least for some of the participating libraries such as Cornell. The initiatives began at a time when we are actively exploring our programs in light of developments such as Google's search engine for information discovery and a growing focus on cyberinfrastructure and the systems that support data-intensive initiatives. There is also increasing pressure to focus digital preservation efforts on the unpublished and born-digital information domain, where preservation concerns are most urgent.

Although research and practice indicate that users increasingly prefer digital information and services, academic and research libraries remain under pressure to continue traditional services too. It is rare to hear about a service being eliminated in order to shift funds into a newly growing area. But the costs of processing and archiving new digital material may cause a significant shift in how funds are distributed among services at many libraries. It is important to try to articulate a preservation program for digital books within the broader scope of library activities and mid-term strategies. Also critical is to envision digital preservation and enduring access by taking into consideration evolving scholarly needs and various information genres and formats.

Acknowledgments

We would like to thank the members of the LSDI team in planning and implementing the digitization project at Cornell. The technical aspects of the project have greatly benefited from vision and hard work of Danielle Mericle, Adam Smith, Marty Kurth, Jon Corson-Rikert, John Fereira, and Frances Webb.

References

The following report provides an overview of challenges faced in large-scale digitization of library materials: Oya Y. Rieger. *Preservation in the Age of Large-Scale Digitization*. Washington, DC: Council on Library and Information Resources, 2008, http://www.clir.org/pubs/abstract/pub141abst.html

Cornell University Library, Digital Preservation Policy Framework, 2006. http://hdl.handle.net/1813/11230

Anne R. Kenney and Nancy Y. McGovern, "The Five Organizational Stages of Digital Preservation," in Digital Libraries: A Vision for the Twenty First Century, a festschrift to honor Wendy Lougee, 2003.

Section 108 Study Group Report, March 2007, http://www.section108.gov/docs/Sec108StudyGroupReport.pdf

Buonora, P. and Liberati, F. (2008). A Format for Digital Preservation of Images: A Study on JPEG 2000 File

Robustness. D-Lib Magazine, 14(7/8) http://www.dlib.org/dlib/july08/buonora/07buonora.html

aDORe: http://african.lanl.gov/aDORe

Fedora Commons: http://www.fedora-commons.org

PREMIS: http://www.loc.gov/standards/premis.

Z39.87: Data Dictionary—Technical Metadata for Digital Still Images. Available at http://www.niso.org/standards/standard_detail.cfm?std_id=731.

Metadata-extraction tools such as JHOVE and NLNZ Metadata Extractor Tool generate standardized metadata that is compliant with PREMIS and Z39.87.

Lorcan Dempsey. 2006. “Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age.” *D-Lib Magazine* 12(4). Available at http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

Book Search Winding Down, May 23, 2008, http://blogs.msdn.com/livesearch/archive/2008/05/23/book-search-winding-down.aspx

According to a study at the University of Denver, most of the problems people perceive with electronic books are related to the difficulty of reading large amounts of text on the screen. Michael Levine-Clark. 2006. “Electronic Book Usage: A Survey at the University of Denver.” *Libraries and the Academy* 6(3): 285-299.