Digitization in the Real World

Lessons Learned from Small and Medium-Sized Digitization Projects

Edited by
Kwong Bor Ng & Jason Kucsma

Metropolitan New York Library Council

Published in the United States of America by
Metropolitan New York Library Council
57 East 11th Street, 4th floor
New York, NY 10003-4605
p: (212) 228-2320 f: (212) 228-2598
Web site: http://www.metro.org

ISBN: 978-0-615-379998-2

Cover Design: Jason Kucsma *(illustration by Smartone Design, licensed via iStockphoto.com)*

Reviewers Committee: Mark F. Anderson, Jill Annitto, Anna Craft, Jody DeRidder, Renate Evers, Wei Fang, Maureen M. Knapp, Sue Kunda, Mandy Mastrovita, Ken Middleton, Emily Pfotenhauer, Mark Phillipson, Alice Platt, Mary Z. Rose, Stacy Schiff, Jennifer Weintraub, Andrew Weiss.

The views expressed in this book are those of the authors, but not necessarily those of the publisher.

Entering the Digitization Universe: One Catalog Librarian's Experience at an Academic Library

Mary Rose (Southern Illinois University Edwardsville)

Abstract

This chapter describes a catalog librarian's experience with an academic library's digital collection initiative. The author discusses how the library handled technical challenges and established policies and procedures during the process of creating its first digital collection. The effects of external pressures from consortial requirements and organizational change are also discussed. The author describes technical decisions specific to the first project and more general technical issues like customization decisions and decisions about filenaming convention. The processes involved in establishing selection criteria and rights and permissions policies are described. The author also provides a brief overview of three subsequent digital projects. The author concludes by speculating on how the library's digital presence will grow in the future.

Keywords: Academic libraries, Catalog librarians, CONTENTdm, Digitization, Digital collections.

Lovejoy Library at Southern Illinois University Edwardsville (SIUE) entered the universe of searchable digitized collections in 2008. We encountered several issues along the way to completing our seminal project. There were technical challenges to be met, and we had to establish procedures and policies. We also encountered external pressures due to our reliance upon consortial services and as a result of organizational changes at the University. This chapter is a narrative of this experience and a speculation about the future.

Background: the preliminary steps toward establishing a digital initiative and vision

In 2006-2007, Lovejoy Library administration took the first steps toward establishing a digital projects initiative by forming a CONTENTdm committee and acquiring access to CONTENTdm software as a member of the Consortium of Academic and Research Libraries in Illinois (CARLI). The software is installed and maintained on CARLI's server. Two Lovejoy staff members received training in the use of CONTENTdm; however, neither staff member was empowered with a mandate to create a digital collection. The initiative essentially stalled. When I joined the University as the Library's first catalog and metadata librarian in May of 2007, I recognized that getting Lovejoy fully engaged in the creation of digital collections was a main priority of the position. The aforementioned staff members immediately and gratefully handed their CONTENTdm workbooks over to me and notified the consortium that I was now the primary contact for coordinating the Library's use of this software. I had never previously used CONTENTdm but became intimately familiar with it over the course of the next several months. Lacking training or experience, I relied heavily on support services at CARLI to effectively leverage the software. I also took a generic metadata creation workshop and studied Dublin Core.

I quickly became aware that two digitization projects were being spearheaded by two tenured faculty librarians as candidates for our initial digital collection: one somewhat aggressively as a grant project and the other more casually without the impetus of a grant. Being naïve with regard to the politics of the organization, I deferred to others who decided to give precedence to the grant-funded project. The CONTENTdm committee subsequently decided that the Library needed a process for evaluating and prioritizing potential digital collections. Perhaps this was a response to the way in which resources had been committed to the first project because of a schedule driven by external funding. Or perhaps it was the usual librarian caution that any new undertaking will grow to unmanageable proportions if fed too liberally. Perhaps the desire for oversight was motivated by recognition that the shape of our accumulated digital collections over time would define the character of the Library to a significant degree, and whether this was ad hoc or directed was not a matter of chance but of choice. Whatever the reason, a digitization selection subcommittee to the collection management committee was proposed by a tenured library faculty member at the first CONTENTdm committee meeting I attended.

The digitization selection subcommittee became entwined with the Library's vision regarding digital initiatives. The subcommittee's charge was officially established as being the body responsible for receiving and evaluating digitization project proposals and making recommendations to the parent collection management committee regarding acceptance and prioritization of said proposals. The advisory group comprising the subcommittee included all of the library faculty administrators plus the Director of Development (essentially the marketing administrator) and the Director of Academic Computing. The subcommittee was rounded out by the Catalog and Metadata Librarian (me), the Electronic Resources Librarian, the Archivist and Special Collections Librarian (serving as chair), and whichever subject librarian was participating in a specific digitization proposal. The group resolved to create a proposal form to guide proponents in describing the subject, extent, rationale, funding, etc. of their project ideas. Selection would be accomplished by carefully evaluating the relevance of a project to the Library's mission and the advantages a digital platform was expected to provide for the particular included items, such as wider accessibility for heavily used resources, easier use of delicate or cumbersome materials, and improved access to text-rich content through electronic searchability. Selection criteria suggested by the Northeast Document Conservation Center were incorporated into the subcommittee's official position. The Center frames selection around three basic questions (Gertz, 2007):

- *Should* [the materials] be digitized? Is the collection important enough, is there enough audience demand, and can sufficient value be added through digitization to make it worth the cost and effort?
- *May* they be digitized? Does the institution have the intellectual property rights to permit legal creation and dissemination of a digital version?
- *Can* they be digitized? Will digitization achieve the goals of the project, given the physical nature of the materials and their organization, arrangement, and description? Does the institution have the technical infrastructure and expertise to create digital files and make them available to users now and in the future?

Challenges encountered during the first digital project

Our pilot digital collection was the KMOX sheet music digitization project. Lovejoy Library's Music Special Collections includes a gift from KMOX of over 48,000 music titles compiled by the St. Louis-area radio station: the live studio orchestra's complete performing music library. The titles date from the early 1900s. A subset of this collection, identified as being published prior to 1923 and hence in the public domain, became the target digital collection. Academic Computing, an entity under the administration of the Library's dean, scanned the sheet music in color at 600dpi, enlarged 400% during scanning and saved as uncompressed tif files. I began working on the project in earnest in January of 2008, with my first real technical task being to understand CONTENTdm enough to design a structure to showcase the collection effectively. Eventually I settled on a strategy: Each piece of sheet music would be what is known in CONTENTdm terminology as a *compound object*. Metadata would be supplied at the object level, meaning each piece of sheet music would have its own metadata but the individual pages comprising a given title would not be described separately.

The Fine Arts Librarian had obtained grant monies to hire graduate student assistants to help with the project. I trained the graduate assistants how to provide what catalogers consider *descriptive* metadata. This is the metadata that is transcribed from the piece being described. In this case, descriptive metadata included the song title, first line of the refrain, and publication information. I also showed the students how to search the Library of Congress's free online authority file (*Library of Congress*, 2009) for authorized forms of names for the lyricists, composers, arrangers, performers, and/or illustrators credited on the pieces. Finally, I created standard notes for the student assistants to apply, such as "piano, vocal" for an instrumentation note, "One color (purple)" describing the cover art, and "Includes advertisements" as a miscellaneous note. I reviewed their work and completed the metadata with subject analysis, detailed cover art description, and additional notes.

Learning how to use CONTENTdm required a tremendous amount of time and energy during this first project. The effort was amply rewarded, however, since the functionality provided by the software suited our application perfectly. The software supports batch population of a collection via tab-delimited files. This facilitated collaborative metadata creation, since the graduate student workers could create Excel spreadsheets with preliminary metadata for groups of titles and then pass them on to me to complete. I subsequently converted the spreadsheets into tab-delimited files and uploaded the metadata into CONTENTdm along with the corresponding images. The compound object structure, in which several images comprise one digital entity, elegantly matches the character of multi-page sheet music. The software also provides the means for creating index boxes, which enhance access to the content beyond full text searches of the metadata. We decided to use this functionality to create index boxes for composers, lyricists, and subjects for this project.

As stated previously, Lovejoy Library's digital collections are created under the consortial umbrella, using CARLI's CONTENTdm server. CARLI's collection of member libraries' digital collections is OAI-harvestable, and CARLI provides a means for member libraries to obtain usage statistics. But with these advantages come some constraints. CARLI requires all of their hosted collections to contain certain metadata fields, including (among others) *Rights* and *Language* fields mapped to the corresponding Dublin Core elements and a *Collection* field mapped to *Relation*. The *Rights* field requirement motivated our CONTENTdm committee to address the thorny issues of intellectual property more promptly than we might perhaps have otherwise; as it was we needed to formulate a policy before publishing the KMOX collection. This proved to be the committee's most important task. The consortium specified that the *Rights* field should identify the intellectual property rights status of the digital resources in the collection and provide direction for users to contact the owner. This field could also be used to inform users of fair use laws. The committee consulted with the University's legal counsel to develop a rights and permissions policy in conformance with these guidelines. The digital rights and permissions statement that eventually evolved through the committee's deliberations authorizes "fair use" of the digital resources, provides references describing the legal limits of fair use, specifies the form of attribution, and provides the means for applying for additional permissions (Lovejoy Library, 2009).

SIUE is responsible for the resources comprising our digital collections. Before we begin a digital collection project, we need to establish our right to create these component digital resources. This can be accomplished by using source materials in the public domain, securing permission for digitization and publication from the owner of the source materials, or actually purchasing the right to digitize and publish source materials. However, the rights status of the source materials is not always unambiguous. For the KMOX sheet music project, items within the public domain were identified as such by having a copyright or publication date prior to 1923 printed on the item. However, as I completed the metadata I noticed that some of the covers exhibited images clearly indicating that they were created after that date. For instance, the cover of "Come to the Fair" featured a photograph of the Trylon, Perisphere, and Helicline at the 1939 New York World's Fair, despite the fact that the music bore a copyright date of 1917. The cover for "The World Is Waiting for the Sunrise" depicted singer Mary Ford, who was born in 1924. Other pieces of music included advertisements for songs displaying later copyright dates. The Fine Arts Librarian consulted with legal counsel about the status of these items. It was decided that we could include these pieces of sheet music in the collection if we didn't provide access to the individual pages that were not in the public domain, an approach that wouldn't affect the usability of the music itself.

Another consequence of using CARLI's server is that our digital collections are subject to CARLI's "look and feel" requirements for uniformity. The consortium allows very little flexibility, as it wants to maintain a consistent look between the collections of member organizations. For our first project, this was actually a blessing. Designing a branded image is a lengthy process requiring resources (graphic design talent and technological tools and adroitness) and research (complying with the look and feel requirements of the SIUE website as a whole). As it was, the decisions I presented before the CONTENTdm committee were straightforward and simple. I made some mockups featuring the school colors in various combinations in the permitted areas; the voting process was fairly painless. I worked with Academic Computing personnel to get an official logo that conformed to CARLI's size constraints with the exact color specified by SIUE marketing guidelines.

Challenges encountered during subsequent digital projects

My second digital collection experience, the digitized presentation of a Civil War diary, was achieved in collaboration with the Social Sciences Librarian and a temporary staff worker under her supervision who had transcribed the entire diary. I learned how to use the transcription function in CONTENTdm and worked with the staff worker to render the transcription she had created into files that CONTENTdm could manipulate, i.e. individual text files with file names matching the corresponding image files.

In the spring 2009 semester, I was the instructor of record for a student's Senior Project course. The student, who had worked in a library for several years and was considering going to library school after completing his bachelor's degree, wanted to learn about Dublin Core metadata and digital collections. Together we designed a project for him to create a Civil War collection under my supervision using digitized letters and ephemera loaned to the Library by an emeritus professor of history. After completing background readings and papers, the student spent about five hours a week at the Library. He collaborated with me to make metadata decisions and learned how to use the CONTENTdm software, successfully completing the project in a semester's time. Working on this collection revealed a shortcoming of CONTENTdm in the way it supports managing metadata for disparate types of materials within a single collection. This project, which my student ultimately named the American Civil War Collection, is comprised of three different types of digital entities: letters, military orders, and songsheets. Adequately describing all three required a total of 27 different metadata fields. CONTENTdm does not have the functionality to organize metadata separately into subsets determined by their relevancy to particular included objects. Metadata manipulation (mapping, editing, etc.) after uploading is performed in a single interface in which all the metadata fields are displayed together. Fortunately in our case the small overall size of the collection meant coping with this limitation wasn't prohibitively awkward.

The same spring the library administration hired a second catalog and metadata librarian, and together we began work on a fourth digital collection. This project featured digital photographs of architectural artifacts designed by architect Louis H. Sullivan and owned by SIUE, accompanied by digitized historic photos of the buildings on which the ornaments originally appeared. We worked with the Fine Arts Librarian and her graduate assistant to plan the organization and presentation of the images and identify the metadata we wanted to include. The graduate assistant gathered the raw metadata which my colleague and I translated into controlled vocabularies. We used the Getty Art and Architectural Thesaurus (AAT) (Getty, n.d.) for terminology for the ornaments themselves, materials of construction, and types of buildings of origin, supplementing the latter with Library of Congress Subject Headings when we felt it would be helpful. We began populating the digital collection in April, a process that took four months due primarily to delays in obtaining some of the images and associated descriptions. Leveraging CONTENTdm to create a meaningful structure for objects in this collection proved challenging. We ultimately decided upon what CONTENTdm calls a *monograph* structure. A CONTENTdm monograph is a compound object with hierarchical levels, analogous to chapters in a book. We organized each of our digital entities to have two subsets (chapters) of images: artifact images and building images. Users click on one of these headings to reveal the images in the next hierarchical level. Although this structure isn't inherently intuitive, we felt it was the best fit from among the options available in the CONTENTdm software. The structure works well when browsing the collection as a whole or via the index boxes we supplied for artifact type and building of origin, but we are less enthusiastic with how it translates into retrieval from keyword searches. CONTENTdm has options for customizing the retrieval display that address some of our concerns, but the fact that the software isolates document- and page-level metadata in the search and display customization functionalities prohibits us from achieving our ideal result.

In summary, the four digital collections that Lovejoy has created to date using CONTENTdm are:

1. KMOX Popular Sheet Music , comprised of 118 objects and 558 jpg files.
2. William R. Townsend Civil War Diary, comprised of 14 objects and 356 jpg files.
3. American Civil War Collection, comprised of 9 objects and 40 jpg files.
4. Louis H. Sullivan Ornaments, comprised of 64 objects and 191 jpg files.

The completion of each project was marked by announcing its availability to the University community and adding a link to the library website. I also created a catalog record in OCLC for each collection, and added all four collections to the CONTENTdm Collection of Collections database (*CONTENTdm*, n.d.).

Issues to be addressed in future projects

Organizational change has provided a source of external pressure concurrent with and affecting the progress of our digital initiatives and priorities. Lovejoy Library's dean left near the end of 2007 after a long tenure as both Dean of Library and Information Services and Associate Vice Chancellor for Information Technology. Academic Computing had reported to the Dean in his latter capacity. Upon the Dean's retirement, the Provost decided to change the organizational structure so that the new library Dean would not have this dual responsibility. Academic Computing merged with the Office of Information Technology Services and now shares with it a new reporting structure separate from library administration. The new system began in July 2008. The interim Director of Technical Services began exploring a team approach to digitization. His plan centered around two major initiatives: purchasing a large format scanner for the library and hiring a digital imaging specialist, which were accomplished in 2008-2009. Library digitization projects could consequently be created without relying on Academic Computing personnel to scan materials. However, all of the aforementioned projects were digitized by various people before the purchase of the library's large format scanner and subsequent hiring of the Digital Imaging Specialist. The team approach has not yet been developed for producing CONTENTdm digital collections.

In fact, creation of the image files began prior to my involvement with each of the projects except the first. As a result, a filenaming convention was never established. We discussed filenaming for the first project, the KMOX sheet music collection. We decided to use a transparent method: The images were named using a combination of the song title, composer name, date, and page number. An example is ByTheLight_Edwards, Gus_1909_001.jpg. This approach doesn't support generalization to future projects. As I researched the issue further, I grew to prefer a more systematic approach to filenaming. This idea inspired me to create an *Image ID* field in the metadata for each sheet music title in the collection, which I populated with an alphanumeric collection-specific accession number. But the actual file names corresponding to the jpg files weren't included in the final metadata: an inadvertent oversight resulting from my inexperience with how CONTENTdm handles tab-delimited files. The problems with file names persisted for all four of the projects described previously in this chapter. The Digital Imaging Specialist is working with me and the other catalog and metadata librarian to establish a convention that works with our scanning equipment defaults. We have decided to adopt a cross-collection systemized convention similar to that described in the *Wisconsin Heritage Online Digital Imaging Guidelines*:

> File names for digital masters and derivatives need to be established before the scanning process. Systematic file naming helps not only to manage the project, but also ensures system compatibility and interoperability. It is generally recommended to assign an eight-character file name and a three-character extension, e.g. aa000001.xxx. This is sometimes called 8.3. File names should adhere to some general requirements. They should be:
>
> - Unique and consistent
> - Alphanumeric (consist of only letters and numbers)
> - Lowercase
> - Free of spaces and tabs
> - Numbered sequentially using leading zeroes (i.e. 001, 002, 003, not 1, 2, 3)
>
> The files can be named after an original source collection or per project, depending on the needs of the local institution. Up to four letters can represent the project abbreviation or original collection name, e.g. hf for Harrison Forman Collection or sccl for Shawano City-County Library. The remaining digits indicate a unique file number. This is often simply sequential numbers prefaced with leading zeros. For example, digital images from the Harrison Forman collection project are named hf000001, hf000002, etc. (p. 5-6)

I reached the end of my digital collection backlog with the completion of Lovejoy's fourth CONTENTdm-based collection in September 2009. I subsequently met with some of my colleagues to brainstorm ideas for additional digital collections. The result was a fantastic array of proposals employing audio and video files, featuring collaboration with other local institutions, and creating scholarly research products on a digital platform. The proposals were presented to the selection subcommittee. The constituency of the subcommittee had been modified to reflect the organizational changes described previously in this chapter. The Director of Academic Computing was no longer a part of the subcommittee and the new Digital Imaging Specialist had been added. Although the majority of the subcommittee greeted the new project proposals with enthusiasm, the role of the selection subcommittee is currently being reconsidered and thus the project approval process is on hold.

We are planning to purchase a server in cooperation with our IT department. Not only will this relieve severe storage problems during digitization workflow, it will also give us the option to explore creating portal pages to our CARLI digital collections or to host some collections locally. The Digital Imaging Specialist has a graphic design background and is highly interested in exploring creative ways to showcase our collections.

Sorting out the process of green-lighting digital projects and the graphic and technological design of locally-hosted portals and collections will doubtless incur long and passionate discussion. The committee-driven process that is the default for all decisions at Lovejoy is not a painless one. Consensus-seeking, while attractive in theory, is impractical in many ways. But it is the culture of this institution and I suspect the culture of many similar institutions as well. Some issues along the way to realizing our digital initiatives thus far have been thoroughly discussed and resolved with thoughtful regard for the future, and some were hastily addressed with the main goal of overcoming a stalling impediment. Some of the best ideas proved insufficiently nimble to adjust to unforeseen developments. Some of the bad seed sown in the interests of forward motion has yet to bear the anticipated troublesome crop. Regardless, we are moving forward into new kinds of projects with a sharper focus on who we are and how we want to present ourselves.

References

CONTENTdm Collection of Collections. (n.d.). Retrieved December 1, 2009, from http://collections.contentdmdemo.com/

Gertz, Janet. (2007). *Preservation and selection for digitization*. Retrieved December 10, 2009, from http://www.nedcc.org/resources/leaflets/6Reformatting/06PreservationAndSelection.php

The J. Paul Getty Trust. (n.d.). *Art & Architectural Thesaurus Online*. Retrieved December 8, 2009, from http://www.getty.edu/research/conducting_research/vocabularies/aat/

Library of Congress Authorities. (2009). Retrieved December 1, 2009, from http://authorities.loc.gov/

Lovejoy Library. (2009). *Digital rights and permissions*. Retrieved December 1, 2009, from http://www.siue.edu/lovejoylibrary/about/digital_rights_and_permission.shtml

Wisconsin Heritage Online Digital Imaging Guidelines (Version 2) (2009, September). Received December 2, 2009, from Wisconsin Heritage Online Wisconsin Library Services.