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: 978-0-615-37998-2

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.

Digitizing a Newspaper Clippings Collection: a Case Study and Framework for Small-Scale Digital Projects

Maureen M. Knapp (John P. Isché Library, New Orleans)

Abstract

How does a small specialty library establish, develop and maintain in-house digital collections? What are the considerations, challenges, and benefits they experience? This chapter describes one library's experience in turning an aging and inaccessible collection of newspaper clippings into a preserved and searchable online collection, which in turn laid a basis for other digital projects. This chapter also discusses considerations, challenges and opportunities observed during their first foray into creating a digital collection.

Keywords: Clippings, Digital libraries, Digital preservation, Digital projects, Digitization, Electronic preservation, Newspaper clippings file, Newspaper clippings, Press clippings.

Background

The John P Isché library is a mid-sized, urban, academic health sciences library serving six schools of health professions at the LSU Health Sciences Center (LSUHSC) in New Orleans, Louisiana. Established in 1931, the library has collected newspaper clippings related to the history and accomplishments of the health sciences institution since its inception, and even today monitors the local papers for pertinent news items. The "newspaper clippings file," as it

came to be called, is an astounding 70 year snapshot of the development of the health sciences in Louisiana. Over 6,000 clippings trace development of LSUHSC through the twentieth century, including such topics as: the people, places and events associated with the LSU School of Medicine, the growth of health infrastructure in Southeast Louisiana and New Orleans, and the development of 20th century health sciences education in Louisiana.

Digital Collection Origins

In 2002, access and preservation concerns with some of the earliest newspaper clippings encouraged the library to investigate digitization as a possible solution. Access points to the collection were limited. The only online access consisted of a locally-created subject database containing basic citations to newspaper articles from 1985 to present. Users had to search the local database by faculty name or department, and then locate the physical newspaper clippings in filing cabinets by call number. The remaining fifty-odd (1933-1984) years of the collection was indexed in a card catalog, stored in the library's back offices and only accessible to library staff.

Numerous problems plagued the physical collection. The newspaper clippings had been stored in filing cabinets as they were collected, which allowed the typing paper to curl heavily over the course of many years. The newsprint itself showed signs of age: rust marks appeared where staples and paperclips had once connected pages, and gaps in the collection were apparent.

A lack of funding and staffing was another concern. Any efforts towards creating a digital collection would have to be inexpensive and make use of staff and resources the library already possessed.

However, to truly understand to physical condition of the newspaper clippings file, and the challenges that would arise once digitization began, one must understand the collection process of gathering the original newspaper clippings. While no documentation exists, the library postulates that even back to the 1930s, a library member would skim the daily local papers from around Southeast Louisiana for any mention of LSU School of Medicine, and its faculty,

staff or students. Once an article was discovered, it was cut out of the paper, dated, and the name of the paper was noted. The articles were glued to standard 8 ½ by 11 inch typing paper, usually several to a page, somewhat in order by date, and the paper was assigned a numerical call number in the order they were received. Later someone would read the articles, underline named entities pertaining to LSU, and assign a subject heading, which was recorded in a small local card catalog. Finally, the pages of clippings were organized into manila folders by year and placed into filing cabinets until further needed. This entire process continued for 50 years.

So basically, the library had a unique local news collection, spanning the majority of the 20th century, collected and stored under questionable archival methods, with limited access to documents before 1985. In order to increase availability and use of the clippings, the library wrote a grant proposal for a small-scale digitization project to scan the newspaper clippings from 1933-1953, streamline cataloguing, and offer public access to the resource online. The grant proposed using Greenstone digital library software, an open source "suite of software for building and distributing digital library collections" (Greenstone digital library software, 2007), to provide access to the digitized newspaper clippings.

Stops and Starts

Though the grant proposal was rejected, the grant writing process did provide a catalyst for action within the library. The small grant requested $3,000 to purchase a flat-bed scanner, computer and optical character recognition software. Library administration was impressed enough with the grant's digitization plan that they provided funding for a scanner, software and travel to a continuing education class on digital projects in 2003. A library staff member began scanning the clippings. However, the library quickly ran into problems. The Greenstone software would not work properly on their secure intranet, and the library lacked a staff member with enough computer programming experience to install and troubleshoot the software properly. In addition, the image quality of the scanned

newspaper clippings was poor, which was attributed to a faulty scanner that did not produce dark enough images. Finally, copyright concerns made library administration hesitant to post the collection online to the general public.

By the time Hurricane Katrina struck New Orleans in August 2005, access, software and image quality issues had put the library's newspaper clippings digitization project on hold. The library's collection was undamaged from this natural disaster. However, it was moved to remote storage for over half a year and the entire library staff was displaced.

During the ensuing hiatus, library staff took several continuing education classes on digitization. "Digitization Fundamentals," a course offered by the Illinois Digitization Institute at the University of Illinois Urbana-Champaign (University of Illinois Library, 2009), was exceptionally useful, as it provided training in digital projects management, standards and organization, as well as an introduction to Photoshop software.

In 2007, an opportunity opened for the library to join the Louisiana Digital library, the state digital library consortium provided through LOUIS: The Louisiana library network (LOUIS: The Louisiana library network, 2009). The library was able to obtain access to OCLC's CONTENTdm platform, which was previously too expensive, as well as the technical infrastructure and support needed to store and access digital assets.

Consortial membership for digital library services addressed many of the problems faced by the library developing an in-house digital collection. The documentation on the technical and operational requirements for participation in the LOUISiana Digital library proved critical. The consortium's style manual for scanning and cataloguing provided guidelines for selecting collections to digitize, scanning practices, post-scanning image manipulation, project workflows, metadata standards, and quality control. Another practical advantage to consortial membership was LOUIS staff support, which provided advice on imaging standards, basic training on the

CONTENTdm software, and a shoulder to cry on when things went awry.

The library began their second try at developing a digital version of the newspaper clipping file in January 2008. As of December 2009, the library has not only met their original goal of digitizing and indexing over 1600 items in the collection from 1933-1953 (LSUHSC New Orleans library, 2009), but also created several other collections.

Work Flow, Image Manipulation and Standards

The format and organization of the newspaper clippings collection created a challenge in regards to digital manipulation and workflow. In order to achieve indexing of items on an individual level, some information that was included only once on a sheet of several newspaper clippings (for example, the name of the newspaper, the date, and most commonly, the clipping's call number) would have to be added to each individual item. Thus, several steps beyond simple scanning and image processing were included in the workflow.

Here are the workflow and standards for creating digital versions of the Newspaper Clipping File:

1. Following consortium standards for creating digital images for the Louisiana Digital library, the full-page newspaper clipping is scanned on an HP Scanjet 8390 flatbed scanner to create an archival black and white image at 300 dpi, 8-bit grayscale and saved as an uncompressed TIFF file on the library server. This creates an archival master version of the original digital image.
2. Using Photoshop, a copy of the archival master version is opened and saved according to file naming conventions for the digital library set forth by the consortium. This creates a duplicate of the archival master that can be manipulated to isolate an individual clipping. This file is the image that will eventually be loaded into the digital collection.
3. The duplicate is cropped to isolate a single newspaper clipping. Pages that have only one clipping on them are also manipulated and cropped to minimize file size.

4. If not visible, the call number, date and newspaper name from the original scan are copied, cut and pasted to the now isolated clipping.
5. Post capture processing is applied. The item is processed for alignment and an unsharp mask filter is applied to correct blurring that might have occurred during the scan process. In addition, the image's histogram is viewed to adjust color intensity.
6. The individual, processed image of the individual newspaper clipping is saved to the server.
7. For pages with more than one newspaper clipping, this process is repeated until all clippings have been isolated.
8. After digital manipulation, the TIFF of the clipping is loaded into the CONTENTdm Project Client. Cursory metadata is entered by a library staff member. The file name, size and location are recorded in a Scanning Log to track progress.
9. The librarian performs Optical Character Recognition (OCR) on the clipping to create an excerpted text field and assigns subject headings. OCR produces an abstract of the first 50 words of the article, which is keyword searchable in the digital library. This takes a bit of time, but it is a good way to review the article and assign the proper subject heading. After a final quality check, the item is approved and uploaded to the digital library. Upon upload, CONTENTdm converts the full resolution TIFF file to JPEG, which is what end-users access when viewing the collection online.
10. CONTENTdm also offers an Archival File Manager, which automatically archives collections in a location specified on our library server as they are uploaded to the online collection. Once a volume is full, it is burnt to an archival quality CD recordable disc, as well as saved on the server.

Cataloging and Metadata

The LOUIS consortium requires collections in the Louisiana Digital library to use the Dublin Core 15 metadata element set (Dublin Core Metadata Initiative, 2008), in addition to non-Dublin core structural

and administrative metadata. CONTENTdm allows up to 125 fields per collection. The library decided to add 3 more metadata fields to the newspaper clippings collection: Call number (to locate the item in the physical files), Full Text (for excerpted text) and Contact Information (so users can contact the library). The following lists the metadata fields used in the newspaper clipping collection.

Field Name (in CONTENTdm)	Type of metadata	Metadata Content	Added by
Title	DC	Title of newspaper clipping	LS
Contact Information	A	Contact information for library	T
Creator	DC	Author of clipping	L
Contributors	DC	Contributor to clipping (rarely used)	L
Subject	DC	Institutional controlled vocabulary, MeSH	L
Call Number	D	Call number for the original clipping	LS
Description	DC	“Newspaper clipping”	T
Notes	D	More descriptive information about content of original clipping, if needed	L
Publisher	DC	Newspaper title	L
Date	DC	Date of publication	L
Type	DC	“Text”	T
Format	DC	“TIFF”	T
Identifier	DC	Mandatory field directs users to identifier URL	T
Source	DC	Library name and homepage URL	T
Language	DC	“En.”	T
Relation	DC	URL to homepage of Newspaper Clippings Collection	T
Coverage – Spatial	DC	“New Orleans (La.)”	T

Field Name (in CONTENTdm)	Type of metadata	Metadata Content	Added by
Coverage – Temporal	DC	Year of publication	L
Rights	DC	Copyright information	T
Cataloger	D	Initials of librarian	L
Cataloged Date	D	Date of cataloging	L
Object File Name	D	File name of item	LS
Image Resolution (Archival)	A	Dots-per-inch of scanned TIFF i.e.: "300dpi"	T
Image Bit-Depth (Archival)	A	"8-bit"	T
Color Mode (Archival)	A	Grayscale	T
Extent (Archival)	A	Pixel dimensions of image (WWWW:HHHH)	LS
Image Manipulation (Archival)	A	"Crop, alignment, unsharp mask, histogram"	T
File Size (Archival)	A	Size of TIFF image in KB	LS
Hardware / Software (Archival)	A	"HP Scanjet 8390, Photoshop, ABBYY FineReader"	T
Digitized By	A	Initials of library staff member	LS
Digitized Date	A	Date of digitization	LS
Full Text	D	Abstracted content from OCR	L

List of metadata elements used in cataloging items. Meaning of symbols: A is administrative; D is descriptive; DC is Dublin Core 15; LS is added by Library Staff, L is added by Librarian, and T is added by Template.

Many of these fields are inserted automatically via a template in CONTENTdm. The remaining fields are divided among project members. The most tedious data entry was entering the Extent and File Size fields for each item. Each clipping's dimension and size is

different, so library staff tends to write these down on a notepad as they scan images for entry, then record them in CONTENTdm and the scanning log later.

Another feature of Content DM is the ability to build a customized controlled vocabulary for the Subject field. This worked to the library's advantage, as the newspaper clipping file possessed a card catalog of subjects. The library uses the newspaper clippings card catalog as a basis to build an institutional controlled vocabulary in the digital library. The card catalog also serves as a reference point to verify names and spellings of affiliated persons. This institutional controlled vocabulary can be shared across digital collections, which is an advantage for future projects related to our institution.

The library soon recognized that other subjects would be necessary to adequately describe the digitized newspaper clippings. Original cataloging varied so much over the years that clippings might only include the name of the person or entity mentioned in the article. The library wanted to add more descriptors, so that articles describing conferences, publications, research grants or other common topics were easier to locate. When applicable, the library consults the National Library of Medicine's list of Medical subject headings (MeSH)(U.S. National Library of Medicine, 2009) for appropriate descriptors in the Subject field. For example, the MeSH term "Congresses as Topic" is used when a clipping discusses conferences, or the MeSH term "Publications" when a clipping mentions a new book or journal article published by one of the institution's faculty. Sometimes, MeSH is not useful, especially when discussing local events such as campus expansion or departmental news. In these cases, a subject heading is created and assigned by the librarian. Clippings in the digital collection can be browsed by year, subject, creator or title. Browsing by date is an interesting way to view the development of institutional history. To further open the collection, keyword searching is enabled in the excerpted text field.

Project Considerations

Storage, standards, documentation, training and staffing were all considerations for this project.

Storage was a huge concern. The deteriorating condition of older newspaper clippings made it evident that storing the physical newspaper clippings in filing cabinets was not conducive to preservation. To address the curling paper, books were used to weight down the paper for several weeks. This did not entirely fix the issue of curling paper, but it did help a little in preparing the clippings for a move to flat storage. After flattening, the files were transferred to acid-free archival folders and placed in clamshell archival storage boxes. Finally, the clamshell boxes of physical files were relocated to the library's humidity controlled Rare Books Room, in order to protect them from humidity and sunlight.

Likewise, the library was heedful of digital storage and the "digital mortgage": how will the library address transfer of archival TIFF files to new formats as software and hardware change? Though the library has yet to encounter a change in image format standards, they did attempt to prepare for this inevitability by storing the collection of archival images in multiple locations, as well as on multiple formats. Having multiple copies also addresses the possibility that some files might eventually become corrupted. TIFF versions of the images are burnt to an archive quality, professional grade CD recordable discs, as well as copied to a location on the library server, which is maintained by our institution and backed up daily to tape at a remote location. This is in addition to the processed JPG file that is available to the public on the Louisiana Digital library. A TIFF of the raw scan of the original newspaper clipping is also retained on the library server.

With multiple storage locations and a complicated workflow, documentation and staff training are also important concerns. The library's consortial membership provided a style manual for scanning, cataloging/metadata standards, and basic workflow suggestions. The library used this as a basis for creating a local workflow policy, which includes detailed directions on image scanning and manipulation as well as step by step directions on how to process the item in

CONTENTdm. A scanning log is used to track size and progress of a collection. The scanning log is simply an Excel file which records the file name, file size, and date of digitization, as well as locations to which the file has been saved.

Regarding training, the library realized it was critical that everyone involved with the project learn Photoshop. The LOUIS consortium takes a 'train the trainer' approach to CONTENTdm, so the librarian was responsible for training local staff on the software after initial training.

This project is staffed with one librarian and two library staff members, who devote about 10 hours a week to this project. Library staff is requested to scan and process 60 clippings per week. Scheduling issues quickly became apparent for the librarian project manager, who has bibliographic instruction and reference desk duties in addition to overseeing digital projects. A supervisor suggested setting aside one day a week to solely devote to digital projects. Friday has since become "Digitization Day" and has worked well in keeping the load of items to be processed and approved by the librarian to a reasonable amount.

Benefits and Challenges

One of the first challenges was software sustainability. The free Greenstone digital library software did not work within the institutional intranet and required higher level technical skills than the library possessed. In addition, problems with the original project scanner resulted in poor quality images that had to be redone.

Support from your institution from inception is critical. Administration has to be on board to provide funding and act as a liaison to other resources, for example, consulting with your institution's legal department about copyright questions. Support from information technology (IT) is also important. Getting our IT department to provide support for open source library software was a challenge that soon put the library's original plans to use Greenstone digital library software on hiatus. One of the benefits of membership in a state digital library consortium is that technical support is

provided in an automated timely manner. In addition, the consortium has direct contacts with the software developers at CONTENTdm, so software concerns are quickly addressed.

The newspaper clippings collection is unique in that it collects clippings from many regional news sources. All materials were published after 1923. Therefore, the work may be protected by copyright until 2018. Violation of copyright was a large concern, so the library decided to restrict access to the images within the newspaper clippings collection to the institutional IP address. In order to share the collection with a larger audience, the collection's metadata is searchable and viewable to anyone. This way, any user can find items in the newspaper clippings collection, and if they are not from the institution, the library works with them to get the information or clippings they need.

Funding is a final challenge. Consortial membership to the digital library is about $2000 a year, while hardware and software ran about $1500 in startup costs. In addition, the library director donated a 21" screen won at a library conference raffle for use with the digital projects computer. Digital imaging is much easier with a larger screen. Grants and scholarships are another source of funding. A scholarship from a regional medical library group helped fund attendance at the first continuing education class on digital imaging and metadata for the librarian project manager. An recent Institute of Museum and Library Services "Connecting to Collections" Bookshelf grant (Institute of Museum and Library Services, 2009) allowed the library to obtain a set of conservation resources and books, which was previously non-existent.

The library now has over 10 years of institutional history available online in a searchable database. Visibility and access to this collection has increased. Indexing though OCLC allows results to appear in Google. As a result, the library has received several inquiries about subjects indexed in the newspaper clipping file from the United States and Italy. The clippings file has also acted as a catalyst for change, inspiring library staff to organize the rare books room, research archival storage methods, and apply for grants. One of the benefits the

library is proudest of is the mentoring opportunity this created. A library staff member who helped start this project recently completed their library degree and went on to become a Digital Initiatives librarian at another local library.

The library has established a workflow and gained experience in digital imaging and management for future projects. Because of the success in creating the newspaper clippings collection, the LSUHSC School of Dentistry started a digital collection of historic photographs. In addition, the library worked with the LSUSHC Registrar's Office to digitize graduation program records, which are now available in a public, searchable collection. Finally, the library is in the planning stages of creating a digitized version of early volumes of the medical school student newspaper. The library also continues to add items to the newspaper clippings collection.

As one can surmise, it has been a long 4 years to produce this digital collection, but once the library established workflow and standards it was much easier to begin other projects. Support from the state library consortium certainly expedited and streamlined the process, and the library recommends state or regional consortium membership to any smaller institution considering developing a digital project. For all the tedious data entry and malfunctioning software, the creation of an enduring, searchable and accessible source of institutional history made the entire project worthwhile.

References

DCMI Usage Board. (2008). *DCMI type vocabulary*. Retrieved December 9, 2009, from http://dublincore.org/documents/dcmi-type-vocabulary/

Dublin Core Metadata Initiative. (2008). *Dublin core metadata element set, version 1.1*. Retrieved December 9, 2009, from http://dublincore.org/documents/dces/

Greenstone digital library software. (2007). Retrieved December 9, 2009, from http://www.greenstone.org/

Institute of Museum and Library Services. (2009). *Connecting to collections: A call to action*. Retrieved December 9, 2009, from http://www.imls.gov/Collections/

LOUIS: The Louisiana library network. (2009). Retrieved December 9, 2009, from http://appl003.lsu.edu/ocsweb/louishome.nsf/

LSUHSC New Orleans Library. (2009). *LSUHSC New Orleans newspaper clippings collection homepage*. Retrieved December 10, 2009, from http://www.louisianadigitallibrary.org/cdm4/index_LSUHSC_NCC.php?CISOROOT=/LSUHSC_NCC

U.S. National Library of Medicine. (2009). *Medical subject headings - home page*. Retrieved December 9, 2009, from http://www.nlm.nih.gov/mesh/meshhome.html

University of Illinois Library. (2009). *Digital services and development -- training*. Retrieved December 9, 2009, from http://images.library.uiuc.edu/projects/newproj.htm