Digitization in the Real World

Lessons Learned from Small and Medium-Sized Digitization Projects

Edited by
Kwong Bor Ng & Jason Kucsma

M 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.

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

Collaborative-Centered Digital Curation: A Case Study at Clemson University Libraries

Emily Gore and Mandy Mastrovita (Clemson University)

Abstract

This article will discuss the authors' experience in building and outfitting a regional scan center to serve Clemson University and the South Carolina Digital Library (SCDL), the state's digital library initiative. The authors describe their experiences regarding the establishment of a new unit armed with the task of providing digital curation, imaging, and technological services within an academic library that previously had very few. A subset of their discussion regarding the overarching observations and challenges will also include issues that have arisen within their multiple imaging production workflows, content management, shared metadata, and preservation responsibilities. Throughout the article, the authors address the pervasive and complex relationships between collaboration, sustainability, storage, preservation and access that they have greeted on a daily basis.

Keywords: Archival materials, Collaboration, Cooperation, Digital images, Digital libraries, Digital preservation, Digitization, Distributed preservation, LOCKSS, MetaArchive, Metadata.

Introduction

Clemson University officially established its digitization initiative in the fall of 2007 by establishing a library unit for Digital Initiatives and hiring a unit head. Most large academic libraries like Clemson established digitization initiatives in the 1990s or early 2000s, but there are distinct advantages to beginning an initiative later. One distinct advantage was that we could learn from others and from the best practices and standards in an already established field. Another advantage is that we could begin to think about the blending of digitization initiatives, institutional repository development, data curation and the preservation of digital assets -- in other words, we began thinking in terms of data curation instead of simply digitization. A third, and possibly the greatest advantage, was that we could join existing collaboratives and be instrumental in starting others. In our opinion, collaboration is the key to building sustainable digital initiatives, so we wanted to make sure we took advantage of collaborative opportunities from the start.

Establishing the initiative

Learning from established best practices, prior experiences and contacts with vendors, Clemson University Libraries decided to equip a scan center and object photography studio as the production center for its new digital initiative. The concept behind the development of this scan center is that it would be used not only for projects centered at Clemson, but also for collaborative projects as part of our statewide digital library effort, the South Carolina Digital Library (http://www.scmemory.org). As one of the 3 core partners for the South Carolina Digital Library (with the University of South Carolina and the College of Charleston), Clemson's goal was to establish a scan center to meet the needs of cultural heritage institutions in the Upstate region of South Carolina. Staffing for these collaborative projects has been covered in part by funding provided by the State Library of South Carolina through Library Services and Technology Act (LSTA) funding. LSTA funding is awarded to states on a formulaic basis by the Institute of Museum and Library Services (IMLS). In addition to part-time staffing, LSTA funds have supported the purchase of one scanning station (Dell computer and Epson 10000XL scanner).

Hiring of Key Positions and Restructuring of Extant Positions

The Digital Initiatives unit began with only a unit head. The unit head identified the need to hire someone to be in charge of digital production as well as a programmer. In addition, one position already existing in the Systems department would become the CONTENTdm specialist since the need for desktop support in the library is decreasing. It was also decided that the Systems unit and the newly formed Digital Initiatives unit would merge under the direction of the Digital Initiatives unit head.

After examining positions in many other digital initiative units throughout the country, it was decided that the digital production position would be filled as a librarian faculty position and that the programmer would be hired as a staff member. The librarian would be sought first and the programmer would follow after equipment and processes were in place. Within several months of advertising the position, the Digital Production Librarian position was filled by Mandy Mastrovita. After Mandy's arrival, students were hired to support the production cycle. Currently, there are six student positions in the unit, with students working up to twenty hours per week. Our grant-funded student works additional hours when his schedule allows. Students come from a variety of backgrounds and majors, but all have in common attention to detail, technological aptitude and a willingness to learn. The programmer position has been more difficult to fill and is currently being re-advertised.

Purchase of Equipment, Installation, and Training

After having reviewed the holdings of Clemson University Libraries, we expected to have to digitize a good deal of maps, manuscript material, photographs, and negative film. Therefore, equipment was selected for purchase based on anticipated scanning needs and the incorporation of digitization best practice guidelines as established by leading institutions such as U.S. National Archives and Records Administration's *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images* (U.S. National Archives and Records Administration, 2004), *JISC Digital Media - Still images, moving images and sound advice* (2010) and IMLS Digital Library Forum's *A Framework of Guidance for Building Good Digital Collections* (IMLS Digital Library Forum, 2008) This included flatbed and large format scanning equipment to be purchased with a budget of $100 thousand dollars. While we anticipated some variation in the kinds of materials that we would receive as a regional scanning center, we did not expect our partner institutions and donors to have as much bound material as they have in their holdings, particularly oversized ledgers, scrapbooks, and yearbooks. Because of this, we may need to purchase a dedicated book scanner in the future or look to work with outsourcing vendors to digitize this material.

Small and Medium Format Imaging: Flatbed scanners

We have two flatbed scanners that can handle both transparencies and reflective media: a Kodak iQsmart3 scanner, and an Epson Expression 10000 XL, fitted with a transparency unit. The Kodak scanner is a higher-grade professional scanner that captures images at a much higher resolution, 5500 dpi. We chose the iQsmart3 nearly two years ago because of its reputation as a professional image scanner; it is one of the few scanners that was designed to accommodate the scanning of glass plates in addition to standard transparency and reflective media sizes. The iQsmart3 was well-represented in prominent libraries and projects at the National Library of Australia, the Wellcome Library, Iowa State University, Yale University, and Brigham Young University. Kodak software, support documentation and service, however, have fallen short. We have found that communication with professional listservs (such as the IMAGELIB-L listserv) has helped bring us in contact with other professional users who have helped resolve some of those issues.

The iQsmart3 was designed as a professional printer's scanner. Sometimes, professional print scanning solutions do not apply to the archival environment. In a photolithographic printing environment (the primary market for the iQ3), the film would be held down on the scanner bed with photographic oil. This optimizes contact with the plate glass and minimizes film flaws such as scratches, which disappear when oiled. Another option would be to tape the film down to the glass. Oil and adhesives are not acceptable when working within archival best practices. We have compromised by laying down residue-free gaffer's tape (adhesive side facing the glass, NOT the film), lining the film up with the mask windows, laying the negatives on the glass without adhesives, then laying the masks on top of the film. Other iQsmart3 practitioners have developed physical (temporary) modifications for project-specific demands.

Our second flatbed scanner is an Epson Expression 10000XL. Although the Epson does not yield such a high dpi count (scanning at 2400 dpi), its interface is simple to understand, and we find that because it looks and acts more like a consumer-level scanner that it is the simplest to learn how to operate. We start our newest trainees on this piece of equipment. Relative to our other scanners, the Epson comes in at a very reasonable price of around $3,000 dollars, for the Photo Edition, which includes a transparency hood unit. The Photo Edition comes with its share of cartridges, cut for 35mm negatives, 35mm mounted slides, medium format film, and large format film. The Expression will scan up to tabloid-size transparent or reflective media (27.94 cm x 43.18 cm), and is a real workhorse.

Large Format Imaging: Digital Scanback

After having established our equipment for small and medium format digitization, we determined our budget (approximately $100 thousand dollars) and room dimensions (6.096 m long x 4.88 m wide x 2.37 m high). Our low ceilings limited us to shooting material from our walls. We decided to work with Academic Imaging Associates (AIA), a valueadded reseller (VAR) to help select a scanning unit that would be the best physical fit for our work environment. Working with AIA, we ultimately decided to go with a Better Light Super 8K-HS scan back and a TTI vacuum easel. The Super 8K-HS scan back unit is a capture device shaped like a large format film cartridge. It slips into the bottom of our large format camera (a TTI Digiflex 45ei), where it is tethered by a cable to a Mac Pro workstation. The scan back has a sensor that reads red, green and blue pixels, and is controlled by a motor that glides the sensor across the back of the camera to capture the image. The Mac workstation runs Better Light's ViewFinder software for focus, tone and exposure control and Adobe Photoshop CS4 for any further image review and optimization. (Collette, n.d.)

Our Better Light scan back system selection also necessitated the purchase of appropriate components and equipment from many smaller sub-vendors. The role of the VAR in articulating specifications and orchestrating the shipment of so much expensive and complex equipment was very important. It required trained professionals who have worked closely with these sub-vendors and built out the equipment themselves. The VAR staff's understanding of where traditional analog and digital photography are distinct and where the lines cross was also important, particularly with scanback units such as ours which are built to fit inside of large format analog cameras. They also supplied us with a trainer who was not only a professional commercial photographer, but had also built out many scanback units like ours. He was able to make adjustments, order additional pieces, and train our staff to work with the camera.

Although we had to arrange the installation of the vacuum unit, the remainder of the building and training was handled by the VAR. The Digital Production Librarian had received previous training and worked with a Better Light scan back system, but there were substantial variations in these two customized Better Light configurations. There were no identical pieces of equipment in the new setup, (new lights, tripods, vacuum easels, light tents, etc.) which made training an essential requirement, as she would be responsible for training staff and students. Training, retraining and refreshment of basic large format techniques were a huge help, and fostered confidence in our unit’s staff and students who would be responsible for operating the Better Light scan back equipment.

Digital Imaging and Analog Tools

We were sure to reserve enough in our budget for this important part of our workflow. Although the use of analog cameras is on the wane, in digitization work, there are many digital-analog hybrids, such as our Better Light scanback, as well as analog tools that are still required for handling photographic materials. There is also the matter of handling of analog film-based media throughout the process of digitization. The extent to which materials that are to be digitized are handled or curated is determined locally.

The generation gap is widening with our students who were born in the late 1980s-early 1990s; they may be familiar with photographic prints, but require an introduction to analog film, negatives, and wet film processes. After we instruct our students to keep abreast of best practices by taking the Cornell University Library's *Moving Theory Into Practice Digital Imaging Tutorial* (Cornell University Library, 2003), we have to spend a good amount of time training our students about halftones, how film grain still needs to be checked with a loupe, that transparencies need to be placed on daylight-balanced light tables, and so on, so that they understand that image quality is not always determined by digital equipment or software settings. Teaching them to check analog film for irregularities throughout the process of digitization is an important part of our ongoing training processes.

Building and implementing workflows

We are thankful to have partnerships with cultural heritage institutions that have yielded a substantial amount of original material for us to work with; these relationships have been cultivated as an earlier part of the digitization process. Our newest challenges lie in building workflows that leverage the skills of faculty, staff and students in an effective way, especially in times when resources are tight. A need for more staff is perhaps the most common complaint of digitization practitioners, and has recently been recorded in the results from UNLV's library digitization survey report as the greatest challenge of its survey participants (Lampert & Vaughan, 2009).

Students and Staff

We have addressed staffing challenges in a way that is similar to digitization programs at peer institutions: we have combined resources and balanced the distribution of work between new and existing staff and units in the library. Digital imaging and metadata creation has been distributed amongst repurposed staff, faculty and students in Special Collections, IT, and Cataloging (Boock, 2008). The staff and faculty dedicate a portion of their time to work on our projects, so, as we begin, we really are only working with a fraction of time spent on our digitization projects. We have depended heavily upon student labor, and have structured many production tasks and responsibilities so that they can be fulfilled and monitored by students. Our senior students, for example, assist by performing quality control on the work created by junior students, followed up with further quality control work by staff and faculty. Some of our Upstate project partners in the South Carolina Digital Library have worked collaboratively with us in metadata creation, but they are often content experts, not metadata experts. Our operations and projects are still new, and these relationships will be negotiated continually over time; we hope we can dedicate more staff to the process of entering metadata, performing quality control, preparing content for upload into CONTENTdm and preserving our images and data.

New Materials and New Workflow Plans

In working with multiple collections, we have learned how to pace ourselves when assimilating new materials in new formats and developing new workflow plans. Thus far, our collections have been relatively small. While starting off with smaller projects is recommended for determining equipment and training benchmarks, it still takes time to develop plans for small collections that vary in dimension and format. Each collection requires the design of different sets of instructions based upon equipment specifications, and the research and analysis of best practices for different media formats (e.g. scanning images from an oversized bound ledger requires different handling and equipment than scanning an envelope full of negative transparencies). The preservation of a consistent workflow is balanced alongside the need to pay extra attention to establishing effective communication of detailed plans and best practices to all students, staff and librarians. Consistency in communication is challenged by having to negotiate the different hills and valleys in everyones' levels of technical training and best practice comprehension, a common problem when working with a diverse group of students, staff and faculty in different areas of specialization (Gueguen & Hanlon, 2009).

File Access and Preservation

Access and preservation are considered throughout the process of digitization. We have trained our students to capture and store master TIFF files locally on our production computers, each is backed up with a RAIDed data storage unit that protects all short-term work. Master images are then cropped and de-skewed, and color profiles are assigned. Students embed pertinent collection and keyword information to master TIFFs using Adobe PhotoXMP. When this is done, NISO Z39.87 technical metadata is extracted using the JHOVE API (*JHOVE - JSTOR/Harvard Object Validation Environment,* 2009). JPG derivatives are generated and moved to the CONTENTdm production server; master TIFF images are copied to our SAN (Storage Area Network) for long-term storage. If working collaboratively with partners at a geographic distance, JPGs are often stored where the partners can access them via the Web for metadata generation and quality control.

Descriptive Metadata

A descriptive metadata worksheet, designed on basic Dublin Core (Dublin Core Metadata Initiative, 2008) and South Carolina Digital

Library elements, is sent as an Excel worksheet to our project partners, or uploaded to GoogleDocs and shared with our student workers and catalogers. The students enter descriptive information about the original items and identify any South Carolina county/region information; this is entered in an unqualified Dublin Core spreadsheet. If the original collection belongs to Clemson, we work with staff in the Special Collections and Cataloging departments to complete the more complex Dublin Core fields (DC.Title, DC.Subject, DC.Description) that require subject and collection analysis to complete the fields appropriately. If the collection belongs to a project partner, we determine which aspects of metadata creation they are capable of handling on their end, and adjust accordingly. All involved work is performed in accordance with best practice guidelines as established in North Carolina Dublin Core guidelines (see North Carolina ECHO, Exploring Cultural Heritage Online, n.d.) and county and region information in accordance with South Carolina Digital Library data fields. When the Dublin Core-based spreadsheet is completed, it is later converted to a tab-delimited text file, uploaded to the CONTENTdm production server for display, and saved on the SAN for ongoing storage.

Web-Based Tools

In the past year, we have developed our production workflow by articulating necessary tasks and procedures, preparing training materials, and identifying bottlenecks. With our staff, students and partners working in different physical areas and requiring ongoing training and access to production data and instructional materials, we have turned to a departmental wiki and Google Docs in our work environment. Using these Web-based tools has alleviated IT networking burdens, facilitated group collaboration, and minimized workstation bottlenecks, especially with spreadsheet data entry. Both tools feature extensive history versioning, which adds an extra layer of security when working with groups; if any mistakes have been made, an earlier iteration of a document can be retrieved with ease.

Our departmental Wiki pages have been easy to update, and have simplified training by minimizing the need to re-explain complex instructions. Thus far, we have placed approximately 100 instructional documents in our wiki, covering topics that include: unit tasks, student scheduling, best practices guidelines, workflow models, training instructions for image capture, metadata entry, and uploading collections into CONTENTdm. We have shared our Web-based documents (training materials, equipment specifications, metadata guidelines, etc.) with our immediate colleagues, and portions of this content with wider networks of digital library professionals at conferences. As of yet, the low cost and accessibility of these tools has outweighed other options. However, we have begun to outgrow this arrangement, and are currently in the process of evaluating more comprehensive project management systems that will help facilitate more sophisticated workflow planning.

Content Management

We have not included a great deal about our CONTENTdm workflow. Our preliminary workflow procedures are more useful to the general reader because they thoroughly address the nature of a collaborative production environment that embodies elements of digital imaging and metadata input with forethought towards preservation. While CONTENTdm functions as a presentation management system, we have determined that over time we will need to manage more complex digital objects and associated metadata than a system such as CONTENTdm can handle. Because of this, we will be moving to an open source system, e.g. Fedora, that supports more robust digital object management and preservation.

Distributed Digital Preservation

By establishing a digital initiatives unit in 2007, Clemson was able to learn from field best practices that establishing a digital preservation plan is part of establishing an initiative. Not only was the establishment of a plan essential but so was the establishment of an infrastructure for preservation. After working with campus IT to secure storage on the SAN for our preservation master files, we investigated digital preservation systems. After identifying existing preservation systems, including OCLC's Digital Archive, SRB/iRODS networks, and LOCKSS-based networks (*HOME-LOCKSS*, 2008), Clemson joined the MetaArchive, a private LOCKSS-based network.

The MetaArchive, a National Digital Information Infrastructure and Preservation Program (NDIIPP) funded project centered at Emory University, utilizes private LOCKSS networks that dynamically replicate and distribute digitized items to multiple file servers in multiple locations (see MetaArchive Cooperative, 2010). Before Clemson joined the network in 2008, the network was established and tested for several years by its original member universities, Emory, Georgia Tech, Florida State, Virginia Tech, Louisville and Auburn. In addition, the network's original partners focused on the preservation of Southern digital culture, and Clemson's material certainly falls into that category. At present, Clemson is working to setup our LOCKSS node in order to place our digitized and born-digital content in the network. Soon the collections we curated over the past year will be replicated and distributed on the MetaArchive network to insure continual access.

References

Boock, M. (2008). Organizing for digitization at Oregon State University: a case study and comparison with ARL libraries. *The Journal of Academic Librarianship 34* (5), 446.

Collette, M. (n.d.). *Scanning backs. How they work.* Retrieved March 10, 2010, from http://www.betterlight.com/how_they_work.html.

Cornell University Library. (2003) *Moving theory into practice digital imaging tutorial.* Retrieved March 31, 2010 from http://www.library.cornell.edu/preservation/tutorial

Dublin Core Metadata Initiative (2008) *DCMI metadata terms.* Retrieved March 31, 2010 from http://dublincore.org/documents/dcmi-terms/

Gueguen, G. , & Hanlon, A. (2009). A collaborative workflow for the digitization of unique materials. *The Journal of Academic Librarianship 35* (5), 470.

HOME-LOCKSS (2008) Retrieved March 31, 2010 from http://lockss.stanford.edu/lockss/Home

IMLS Digital Library Forum (2008) *A framework of guidance for building good digital collections*. Retrieved March 31, 2010 from http://www.niso.org/framework/Framework2.html.

JHOVE - JSTOR/Harvard object validation environment. (2009). Retrieved on March 31, 2010 from http://hul.harvard.edu/jhove/

JISC Digital Media - Still images, moving images and sound advice (2010). Retrieved from http://www.jiscdigitalmedia.ac.uk/

Lampert, C., & Vaughan, J. (2009). Success factors and strategic planning: Rebuilding an academic library digitization program. *Information Technology and Libraries*, *28*, 123.

MetaArchive Cooperative (2010). *About MetaArchive: collaboratively preserving our digital heritage.* Retrieved March 31, 2010 from http://www.metaarchive.org/about

North Carolina ECHO, Exploring Cultural Heritage Online (n.d.) *North Carolina Dublin Core guidelines*. Retrieved March 31, 2010 from http://www.ncecho.org/dig/ncdc.shtml.

U.S. National Archives and Records Administration. (2004). *Technical guidelines for digitizing archival materials for electronic access: creation of production master files – raster images*. Retrieved March 31, 2010 from http://www.archives.gov/preservation/technical/guidelines.html