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.

Developing an Institutional Repository at Southern New Hampshire University: Year One

Alice Platt (Southern New Hampshire University)

Abstract

In 2008, Southern New Hampshire University was awarded a three-year, $500,000 national leadership grant from the Institute of Museum and Library Services to create a digital repository using DSpace open source software. Events from the first year of the repository's development are presented and discussed. Key elements addressed include the challenges involved with customizing the DSpace infrastructure, creating standards for access and master files, implementing metadata standards, and developing digital preservation policies. The value of cross-departmental participation is shown, and the importance of planning for digital preservation is presented.

Keywords: Best practices, Digital library, ETD, Electronic, Interfaces, Institutional repository, Open source, Scans.

Introduction

In 2008, Southern New Hampshire University was awarded a three-year, $500,000 national leadership grant from the Institute of Museum and Library Services (IMLS) to create a digital repository using DSpace open source software. The inspiration for the project was a collection of student theses and dissertations from the School of

Community Economic Development (SCED). SCED is a unique program with participation from all over the world, particularly the United States and Tanzania, and also countries such as Uganda, Peru and the Philippines. Like many thesis collections, the projects were printed using consumer-grade equipment, and only one copy was bound and saved for the library. The international nature of the projects, in addition to the danger of losing them to deterioration, made them an attractive collection for beginning a digital repository. Faculty papers from the International Business program were also included in the grant project, to make papers once only accessible from a professor's office available to the world.

Many institutions lack the financial and human resources to build a successful digitization program. The gap between resources available versus resources required can often be bridged by a grant; a search for IMLS grants from 2004-2008 using the keyword "digitization" shows that at least 57 IMLS grants were provided to libraries and museums for digitization projects (IMLS, 2009). Like most institutions, the repository at SNHU's Shapiro Library could not have come to fruition without grant assistance.

The following pages share the Shapiro Library's experiences during the first year of repository development.

The People Involved

Digitization programs need a strong level of organization and administrative support to succeed. Programs that only live within the walls of the library without buy-in from administration and other departments are at risk of failure for lack of support. The Shapiro Library's digital repository is managed by a Digital Initiatives Librarian, who receives support from the Digital Content Specialist, two graduate assistants, and two cross-departmental committees: the Implementation Committee and the Policy Team.

The Digital Initiatives Librarian is responsible for managing the repository, including creation of metadata standards, scanning workflows, policy development, and quality control. The Digital Content Specialist creates descriptive keywords, and writes abstracts for the theses. Two graduate assistants were hired to execute the scanning, optical character recognition (OCR), and access file creation.

The Implementation Committee was initially organized to prepare the grant application, and after the grant was received, organized the necessary infrastructure. Represented on the committee are the Library Dean, the Electronic Resources Librarian, the University Webmaster, the Dean of the School of Community Economic Development, and both the head of the IT department and the IT programmer committed to the project. The committee hired the Digital Initiatives Librarian and a Digital Content Specialist, who both subsequently joined the committee. The Implementation Committee continues to meet on an as-needed basis to monitor the repository's development.

While some members of the Policy Team are consistent with the Implementation Committee, the focus for this group is to determine policies for the repository and discuss other questions that might arise, whether they are related to file format, collection development, or metadata. Because of the nature of the team, there are more librarians represented: the Electronic Resources Librarian, Technical Services Librarian, and the Access Services Librarian are all part of the team, as well as the Digital Initiatives Librarian, Digital Content Specialist, and the Library Dean. Also on the team are the IT programmer and the Associate Dean of the Faculty. The associate dean's participation is effective in keeping the university administration informed on the progress and policies of the repository. The Policy Team initially met every two weeks, and continues to meet at least once a month.

Developing the Technical Infrastructure

After identifying the initial collections for the repository, the Implementation Committee selected the digital repository software and the hardware on which it would reside. Oya Y. Rieger (2007) explains that when selecting software, a number of factors should be considered, including matching your institution's needs to the software's features, considering what resources will be required to install and maintain the software, and assessing the overall usability for both staff and end-users. Often the question might arise: to open source, or not to open source? While using open source software is the current trend, institutions should look closely at their resources to determine if they can support the technological and human resources required to work with open source software packages.

The grant awarded to SNHU included funds to hire the Digital Initiatives Librarian and the Digital Content Specialist. It also financially supported time spent working on the grant by other positions already in place, including IT. Assessment of these resources determined that enough support was available to consider open source software. DSpace stood out as the most widely-used open source institutional repository software package available for academic library use, with an active user community and a wide array of resources available (DuraSpace, 2009, Resources).

The differences between implementing open source versus proprietary software quickly became apparent. While DSpace is advertised to be useful "out of the box," this is not a realistic assertion (DuraSpace, 2009, About DSpace, para. 1). A certain level of programming skill and time is required in order to customize the software. In DSpace, the level of programming needed to make customizations beyond changing the color scheme of the website can be daunting for someone without experience in both programming and website design. The Digital Initiatives Librarian's web design skills and the IT programmer's skills were both needed to make most of the necessary customizations to the user interface.

Community support - While the DSpace community is very active, with a well-populated wiki and listservs for general and technical questions, it is also a complex community. Users vary by what platform they work on (Linux vs. Windows) and what version of DSpace they use. During the time of SNHU's installation, most of the user community was working with either DSpace 1.4 or 1.5. To further complicate things, some users of 1.5 were using what is known as the JSP user interface, while others used the XML user interface – each involving different programming methods for customization. Therefore, not all questions and answers posted by the community are relevant to one's needs. One example encountered was a DSpace wiki entry explaining how to change the DSpace code to enable linking authors in a simple item record. When the code did not function properly, the question was posted to the DSpace tech listserv. Another community member explained that the encoding described in the wiki had changed in version 1.5 (Platt, 2009). Additionally, answers to questions regarding installation varied widely depending on if the user was on Linux or Windows. While DSpace does have a large user community, that community requires some careful navigation.

Professional development – In early June 2009, the NITLE consortium presented a timely DSpace workshop (NITLE, 2009). The variety of sessions provided a strong background to DSpace's capabilities. One session in particular, "Developing Interfaces and Interactivity for DSpace with Manakin Workshop" by Eric Luhrs of Lafayette College, was extremely helpful, providing tools and the know-how necessary to make customizations to the XML user interface (Luhrs, 2009). Without the benefit of this interactive instruction, the learning curve involved would have been much more difficult to transcend.

The experience at the DSpace workshop points to the importance of this type of professional development in the rapidly-changing digital library environment. Conferences such as the Open Repositories Conference and the Joint Conference on Digital Libraries have both included specific DSpace sessions and workshops in the past. A simple search of the web reveals user groups and workshops available for other digital library platforms, including proprietary software such as OCLC's CONTENTdm. Providing funding for librarians and IT staff to attend these types of educational events should be a priority for any institution embarking on a digitization project.

From Paper to Electronic

Creating an electronic record for access involves metadata authoring, scanning, and access file creation.

Metadata – Metadata standards should be determined before the first item is ever added to the repository. Because qualified Dublin Core is installed with DSpace by default, and because Dublin Core is the leading schema for describing digital resources, it was selected for the schema. Determining which elements to make available in the DSpace submission form was more challenging. Not every element should be used to describe a digital object – not all are appropriate for all collections. Besides, the time-consuming nature of metadata entry requires that standards be chosen with efficiency in mind. Michael Boock and Sue Kunda (2009) explain how creating a metadata record for both the DSpace repository and MARC catalog can take up to an hour per record, even when students create the majority of the descriptive metadata (p. 300-302). While it is important to consider descriptive, administrative, structural, and preservation metadata, these elements must be chosen carefully to achieve thorough, but cost-effective item description.

The CDP Metadata Working Group's "Dublin Core Metadata Best Practices", the DCMI Usage Board's "DCMI Metadata Terms", the Scholarly Works Application Profile as described by Julie Allinson (2008), and the Networked Digital Library of Theses and Dissertations' metadata standard (Atkins, Fox, France, & Suleman, 2008) were all examined. From these best practices, 32 qualified Dublin Core elements were selected, with the intention that any item added to the repository could be appropriately described using some or all of these elements. Approximately 20 of these are used to describe the SCED thesis projects in particular.

Scanning – During the development of the DSpace infrastructure, the scanning workflow was also launched. The initial collection of student theses and dissertations from SCED proved to be challenging to scan. Part of the purpose of the SCED thesis project is to document work completed by the student in the field, outside of the classroom. To that end, most of the theses, collected from 1984-

present, include large appendices of documentation including letters, financial statements, marketing materials, photographs, architectural plans, and even a wall calendar used as a fund-raiser. Additionally, students were given the opportunity to be creative in their presentation, often using color, graphs, and decorative fonts.

Sample theses were selected and scanned by the graduate assistants to test how the scanner and OCR software would handle the diverse materials. These initial scans immediately raised questions. There were not yet policies in place for how much information should be captured in the scans, causing uncertainty when incidental color was encountered, such as flyers printed on colored paper. Additionally, there was confusion surrounding the fact that the digitization process includes preserving master files, saved in traditional TIFF format, in addition to the PDF files created for access. The IT staff was not prepared to store and preserve this large collection of master files, and panic arose about their massive size – the files, scanned at 600 dpi in grayscale or color, were 30 to 80 megabytes each. This "megabyte shock" is not unusual, particularly at small institutions; Stacy Nowicki (2008) also noted problems with large TIFF files at Michigan's Kalamazoo College.

After much discussion, the Policy Team agreed to scan the papers for their intellectual content only. Best practices from the California Digital Library (2008) and the CDP Digital Imaging Best Practices Working Group (2008) were consulted to determine digitization standards: a 600 dpi setting for black and white pages, and 500 dpi for grayscale or color pages. This 500 dpi setting resulted in a minimum of 4000 pixels on the largest side of the scan, in accordance with these recommendations (California Digital Library, 3.6.1; CDP, p. 8). It should be noted that if the pages were significantly a different size, the dpi setting would be adjusted to meet this parameter. Master files are saved in TIFF format (California Digital Library, 2008, 3.2). Grayscale and color are only used when necessary to preserve the intellectual content of the document, leaving most of the pages to be scanned as black and white. As a result, the master files are much smaller; the black and white scans are approximately 4 megabytes each.

The solution to the color question requires a certain amount of human judgment, but is viable because the Digital Initiatives Librarian and the graduate assistants conducting the scanning are located in close proximity to one another, facilitating an environment for quick decisions. David Lowe and Michael Bennett (2009) state that the Internet Archive chose to scan all their documents in color, eliminating the need for human judgment (p. 210).

Access File Creation – After creating master TIFF files, it is necessary to convert them for public access. The Portable Document Format (PDF) format, processed so that full-text searching is possible, is ubiquitous among subscription and open access academic databases. It was the obvious solution for our collection.

To enable full-text searching, the TIFs were processed using optical character recognition (OCR) software. ABBYY FineReader 9 Professional was selected, based on a review in PC Magazine (Mendelson, 2008). This feature-rich software enables OCR recognition and error-checking in multiple languages, and performs well with most text, including text printed with a dot-matrix printer, and text formatted in blocks, such as in newsletters and flyers.

From FineReader, the graduate assistants are able to save the PDF with an option called "text under image," saving the corrected OCR text in an invisible, searchable layer under the scanned page image. In order to keep the size of the file reasonable, the PDF images are saved at 300 dpi; to enhance accessibility, the option for creating a tagged PDF is selected (Johnson, 2004).

After the PDF is created, it is opened in Adobe Acrobat, and additional metadata is added to the file's properties, including title, author, and copyright status.

Information in the repository should be not just available, but accessible to all. This includes maintaining file sizes to enable faster load times, ensuring that even users with dial-up modems can download the files in a reasonable amount of time. According to a survey led by John Horrigan at the Pew Internet & American Life Project (2009), seven percent of Internet users in the United States are using dial-up services at home (p. 7). While seven percent sounds small, it is equal to approximately 9 million households in the United States, out of the 129 million counted by the U.S. Census by July 1, 2008. International user statistics vary widely, but it would be best to avoid frustrating any users with unnecessarily large file sizes, thus increasing the viability of the collection.

Therefore, nearly all of the projects are split into two PDFs. Because the bulk of most of the thesis projects is the supporting documents in the appendix, the papers and their appendices are saved as separate PDF files. Of the first 88 student projects scanned, the average file size of the project paper by itself was 2.33 megabytes, with a median of 1.73. The appendices' average was 8.29 megabytes, with a median of 5.2. Both the main paper and the appendix PDFs are available from the same item record in DSpace.

To improve the access files' longevity and accessibility, the PDFs are saved as PDF/A when possible. Roger Reeves and Hans Bärfuss (2009) explain the International Standards Organization's (ISO) goal for PDF/A is that it "provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files" (The Goal of PDF/A). One example of the advantage of PDF/A is ensuring that elements such fonts are embedded in the file, so they display properly even if the user does not have those particular fonts already loaded on his computer. The utility provided in Adobe Acrobat Professional 9 was used to save the files in PDF/A format.

The Often-Missed Point: Digital Preservation

While all of the decisions involved with customizing the DSpace user interface were being addressed, one major component of the digitization program was not addressed: the concept of digital preservation. While it was understood that digital preservation was an issue, it was uncertain how preservation would be accomplished.

The ICPSR Digital Preservation Workshop at the University of Michigan was an excellent opportunity to learn more about digital preservation. This five-day, in-depth workshop made it clear that if an institution presents digital documents online, there is an assumption that they will be preserved there forever – much like a book on the shelf is expected to be readable ten, fifty, or even hundreds of years after it is bound. However, digital files are fragile in their own way, and are susceptible to obsolescence, storage media problems, and other issues (Cornell, 2007, Tutorials, chapter 3: Obsolescence & Physical Threats).

While digital preservation is a complex topic with many components and considerations, the primary concern was how to adequately care for the master TIFF files. Each image must be preserved in the event that the access PDF file becomes corrupted, or when PDF is superseded by a new access file format. The Digital Preservation Tutorial developed by Cornell makes it clear that institutions can not burn files to CDs, put one CD on the shelf, another in someone's garage, and believe they have preserved their files (2007). According to the tutorial, even CD standards have changed over the years, and early formats are now obsolete (Chapter 3: Obsolescence, Chamber of Horrors, Disk Media). It is also apparent that backing up files without including any descriptive information is still not adequate preservation; how many of us have opened a floppy disk and wondered, "What the heck is all this stuff?" Master files must also have their own metadata associated with them to describe what they are. But learning how to preserve these files, as well as adequately preserving the access files and their associated metadata in these early years of digital preservation, is a challenging process that has not been adequately addressed during the first year of repository development at the Shapiro Library. It is probable that many other institutions have also not addressed their own digital preservation questions, or even asked them.

The problems with file obsolescence and data backup are just one small component of creating a digital preservation program. The guidelines presented by the ICPSR workshop are a helpful resource in determining how to ensure that digital preservation at the Shapiro Library is compliant with standards described by the Reference Model for an Open Archival Information System (OAIS), an industry standard. Much progress is anticipated for the second grant year.

Conclusion

The myriad of details involved with creating a digital repository at Southern New Hampshire University were more complex than anticipated. Learning and implementing standards for metadata, master files and access files was time-consuming, but taking the time to establish standards in the beginning doubtless saved a great deal of trouble for the future. Even so, it will be necessary to keep up with developing industry standards, and it would not be surprising if further adjustments are needed down the road. A digital repository is much like a physical building; periodic maintenance, remodeling, and wear and tear should be anticipated and expected.

The Shapiro Library's digitization program has strong administrative support, participation from several university departments, and strong financial resources. The repository will become a successful program for the university long after the grant period concludes.

References

Allinson, J. (2008). Describing scholarly works with Dublin Core: A functional approach. *Library Trends 57*(2), 221-243.

Atkins, A., Fox, E., France, R. & Suleman, H. (2008). *ETD-MS: an interoperability metadata standard for electronic theses and dissertations*, ver. 1.00, rev. 2. Retrieved from http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html

Boock, M., & Kunda, S. (2009). Electronic thesis and dissertation metadata workflow at Oregon State University Libraries. *Cataloging & Classification Quarterly, 47*(3/4), 297-308. doi:10.1080/01639370902737323.

California Digital Library. (2008). *CDL guidelines for digital images.* Retrieved from http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html

CDP Digital Imaging Best Practices Working Group. (2008). *BCR's CDP digital imaging best practices version 2.0. BCR*. Retrieved from http://bcr.org/dps/cdp/best/digital-imaging-bp.pdf

CDP Metadata Working Group. (2006). *Dublin Core metadata best practices: version 2.1.1. BCR.* Retrieved from http://www.bcr.org/dps/cdp/best/dublin-core-bp.pdf

Cornell University Library. (2007). *Digital preservation management: Implementing short-term strategies for long-term problems.* Retrieved from http://www.icpsr.umich.edu/dpm/

DCMI Usage Board. (2008, January 14). *DCMI Metadata terms.* Retrieved from http://dublincore.org/documents/dcmi-terms/

DuraSpace. (2009). *DSpace.* Retrieved from http://dspace.org

Horrigan, J. (2009). *Home broadband adoption 2009. Pew Internet & American Life Project.* Retrieved from http://www.pewinternet.org/Reports/2009/10-Home-Broadband-Adoption-2009.aspx

Institute of Museum and Library Services. (2009). *Search awarded grants.* In Grant Search. Retrieved from http://imls.gov

Johnson, D. (2004). *What is tagged PDF?* In *Accessible PDF Learning Center.* Retrieved from http://www.planetpdf.com/enterprise/ article.asp?ContentID=6067

Lowe, D. B. & Bennett, M. J. (2009). A status report on JPEG 2000 implementation for still images: The UConn survey. *Archiving* 2009, 6, 209-212.

Luhrs, E. (2009). *NIS Camp: Developing interfaces and interactivity for DSpace with Manakin.* Retrieved from http://nitlecamp.pbworks.com/f/manakin-workshop-slides.pdf

Mendelson, E. (2008). *ABBYY FineReader Professional 9.0. PC Magazine.* Retrieved from http://www.pcmag.com/article2/0,2817,2305621,00.asp

NITLE Information Services. (2009). *Nitlecamp.* Retrieved from http://nitlecamp.pbworks.com/

Nowicki, S. (2008). *Using DSpace for institutional repositories.* Retrieved from http://hdl.handle.net/10090/4522

Platt, A. (2009, Oct. 6). *Two theme modification questions. Message and responses,* archived at http://sourceforge.net/mailarchive/forum.php?forum_name=dspace-tech

Reeves, R. and Bärfuss, H. (2009). *PDF/A – A new standard for long-term archiving*. PDF/A Competence Center. Retrieved from http://www.pdfa.org/doku.php? id=pdfa:en:pdfa_whitepaper

Rieger, O.Y. (2007). Select for success: Key principles in assessing repository models. *D-Lib Magazine, 13*(7/8). doi:10.1045/july2007-rieger

U.S. Census Bureau. (2009). *Annual estimates of housing units for the United States and States: April 1, 2000 to July 1, 2008 (HU-EST2008-01)*. In *Housing Units: State Housing Unit Estimates: 2000 to 2008*. Retrieved from http://www.census.gov/popest/housing/HU-EST2008.html