Developing a repository at Southern New Hampshire University: a case study

Alice Platt

Digital Initiatives Librarian

Southern New Hampshire University

Citation: Microform & Imaging Review. Volume 39, Issue 1, Pages 18–23, ISSN (Online) 1865-8458, ISSN (Print) 0949-5770, DOI: 10.1515/mfir.2010.003, /January/2010
Submitted: September 15, 2010. **Published:** December 4, 2010.

Introduction

In September 2008, Southern New Hampshire University was awarded a three-year, $500,000 grant from the U.S. Institute of Museum and Library Services (IMLS) to build a digital repository using DSpace open source software.[1] Creating a digital repository from scratch proved to be a challenging process requiring a great deal of communication and patience. It is anticipated that sharing some of our experiences from the first year of development will help other institutions who are working on their own institutional repository, or just contemplating one.

Getting Started

The SNHU Academic Archive is initially focusing on intellectual output from the School of Community Economic Development (SCED). Student thesis projects from SCED take place in communities throughout the United States as well as countries including Tanzania, Uganda, and the Philippines, resulting in a need for research to be conducted from a distance. Previously, this was nearly impossible; the majority of the SCED projects were limited to one non-circulating spiral-bound hard copy.

Soon after the grant was awarded in September 2008, a 4-terabyte server was purchased, a DSpace 1.5 test bed installed, and high-priority projects from the SCED collection identified. With the assistance of legal counsel, a license agreement was created to secure every alumnus' permission to digitize and distribute their thesis projects worldwide. Approximately 145 of the 800 projects in the library's collection were given priority, and these alumni were initially contacted by mail in January 2009.

The IMLS grant included funding for two positions: one two-year position for a Digital Content Specialist, who generates subject-specific metadata for the documents, and a three-year position for a Digital Initiatives Librarian. The Digital Initiatives Librarian manages the customization of the repository, metadata standards, scanning workflows, and policy development. An existing employee from the Computing Resources department was funded by the grant to implement the DSpace infrastructure and make customizations to the software. The digital initiatives librarian was hired in January, and started work in March; the digital content specialist was subsequently hired and started working in mid-June. Two graduate assistants were also brought on board, primarily for scanning, optical character recognition (OCR) processing, and basic metadata creation.

While some of the hardware was acquired before the digital initiatives librarian came on board, the scanning equipment and software needed to be purchased. The librarian conducted research on the leading available products and concluded that the EPSON 10000XL was the best flatbed scanner on the market for professional output at a reasonable price. The OCR software ABBYY FineReader Professional 9.0 was also selected for its strong industry reviews and reported efficiency with editing and reading unclear type.[2]

An additional dimension to the licensing process was that approximately 200 of the projects were from students participating in SNHU's partnership with the Open University of Tanzania. There was initially concern that students in Tanzania would also need to license their work under a Tanzanian licensing agreement. Fortunately, however, virtually all of these students had already released permission to SNHU via a standard copyright statement included in their student projects.

Metadata

When initially experimenting with how to submit an item in the test bed, it became clear that the default Dublin Core elements available in the DSpace submission form were not entirely appropriate for the items that would be submitted into the repository. Because the elements used in the submission process must be programmed, it was in everyone's best interest to determine all of the elements that might be needed for any resource that might be archived in the future as well as the present.

Leading best practices were consulted to determine the most effective Dublin Core elements, including the CDP Metadata Working Group's "Dublin Core Metadata Best Practices", the Dublin Core Metadata Initiatives' "DCMI Metadata Terms", and the Scholarly Works Application Profile.[3] In addition, the "MARC Value List for Genre Terms" was consulted for the controlled vocabulary choices for the 'type' element. Determining elements was challenging, since at times these best practices actually contradict one another. The librarian was particularly interested in choosing elements with the greatest opportunity for interoperability that would stand the test of time; therefore, very few custom qualifiers were used. The elements were refined after test-cataloging a few documents. A data dictionary was then created to define how each metadata element would be used, and the elements were programmed into DSpace by the Computing Resources staff.

In the end, custom qualifiers had to be created to describe the advisor, committee member, school, program, and degree associated with the documents; these were appended to the 'contributor' and 'description' elements. (See Table 1 for a list of all elements and their definitions). Although Dublin Core guidelines stipulate that no elements are required, it was determined that every resource in the repository should have both a date and a title. These elements were programmed as required for the submission process.

Scanning Begins: Challenges Encountered

The initial materials to be scanned were 25 years' worth of student theses projects completed by students in the School of Community Economic Development graduate program. Upon examination, it was apparent that some challenges would be encountered during the scanning process, and that assessment proved to be an understatement. The quality of the printing varied widely. Many of the older documents were printed using dot-matrix printers or typewriters. Most of the projects were comprised of approximately 20 pages of typewritten text, but the great majority also included a large appendix of letters, photographs, photocopies of reference material, marketing materials, tabs, non-standard sized pages, and color text and charts. One project even included an entire U. S. Government training publication. It was a challenge to create standards for scanning and describing such non-standard objects.

When the scanning began, the graduate students were instructed to scan everything in TIF format, at 600 dpi. The FineReader software recommended grayscale scans, and so all pages that did not include color were scanned grayscale; color was scanned as color, including such items as flyers printed on colored paper. The first week of scanning quickly filled up the server that had been made temporarily available for the project. The TIF files created were much, much larger than anyone had anticipated; one page scanned grayscale at 600 dpi was approximately 30 MB; color pages were approximately 80 MB. While some student projects were only 15-30 pages, others were over 100 pages. A different policy for the master files had to be quickly determined.

Access, or preservation? The question that really needed to be addressed was the distinction between scanning for preservation versus scanning for access. If access files without master file backups were all that was needed, scanning could take place at much lower resolution. However, the purpose of the grant was to digitize for preservation *and* access, following industry best practices. It did not make sense to spend so much effort scanning files merely for access purposes, and it did not follow best practices to do so. The digital librarian agreed to take a closer look at best practices for scanning, to see if there was a way to make the file sizes smaller while maintaining acceptable master files.

The California Digital Library and the CDP both recommend the ppi for grayscale and color scans fall between 4,000-6,000 ppi on the document's longest edge.[4] The files that had been scanned so far in grayscale were more than 6,000 ppi on the longest edge of the document. There was definitely room for adjustment. After some experimentation, it was determined that these pages could be scanned at 500 dpi to fall within those parameters, considerably reducing the file size.

Additionally, experiments were conducted to see if the OCR software could handle black and white scans as accurately as grayscale scans. The graduate assistants reported that it worked just fine, and in fact, in many cases the OCR was actually improved because there were fewer speckles on the pages. This was great news, since a black and white TIF scanned at 600 dpi (as recommended for black and white files by the CDP) was only 4 MB, compared to approximately 30 MB for grayscale.

Once these technical changes were made, there was a great deal of relief that so much growing room was created as a result. However, the nature of the documents also required some policy decisions. Color was a big question. For example, more than one student chose to print the entire project on textured resume paper with a blue marble print. Many more students used color ink when it was not necessary. Even with the new scanning parameters, the resulting files would be prohibitively large. The Policy Committee, tasked with creating policies for the repository, agreed that the purpose of the project was to preserve the *intellectual information* in the documents. Therefore, it was appropriate to only preserve instances of color where the color added intellectual value to the document, such as if it was necessary to understand the information being presented, such as a multicolor pie chart, or for legibility purposes, such as black type in a red box.

Fortunately the graduate assistants took all of these changes in stride and were enthusiastic about creating and implementing the new standards.

Access Files

Once the questions and issues involved in the scanning process were ironed out, the creation of access files had to be determined. In keeping with best practices, all access files would be saved in PDF format. The conversion from FineReader to Acrobat created tagged PDFs that ensured a high level of accessibility; the documents would be converted to PDF/A in Acrobat to ensure long-term access.[5]

The access files were created as 300 dpi PDFs using system fonts only, from the OCR software. The PDF was opened in Adobe Acrobat 9.0 Professional, and the following enhancements were made:

- The Document Properties were updated to include author, title, and the name of the university in the subject field. The tab indicating that the document is under copyright protection was also selected.
- If the document included an appendix (a common aspect to the student projects), the appendix was extracted so there would be one PDF for the main paper, and one for the appendix. This was done to keep the file sizes reasonable enough so that dialup users could access the files with less difficulty.
- The pages of the main files were renumbered so that the page number displayed in Acrobat would be the same as the page number on the document.
- Page numbering within appendices tended to be too erratic for renumbering to be practical.
- The Acrobat Preflight option was then used to save the documents as PDF/A-1b.

Future Challenges

Library staff is expecting to stay busy keeping up with changes in digital preservation. While TIF and PDF were respectively chosen as master and access files because of their reputation for standing the test of time, it is probable that these preferred formats will eventually change. Additionally, updates to the DSpace repository software are certain. Committing to a repository, particularly an open source repository, is a long-term endeavor.

What is not mentioned in this article is the nontechnical challenge of inspiring other schools at SNHU to participate in the repository program. As multitudes of scholarly literature will attest, this will be no small feat, and no repository should be started without making those considerations. It is hoped that the initial collection of student papers from SCED, combined with administrative support from the university, will inspire others to engage their students and faculty to submit to the repository.

Acknowledgement

This project is made possible by a grant from the U.S. Institute of Museum and Library Services. To visit the SNHU Academic Archive, visit http://academicarchive.snhu.edu.

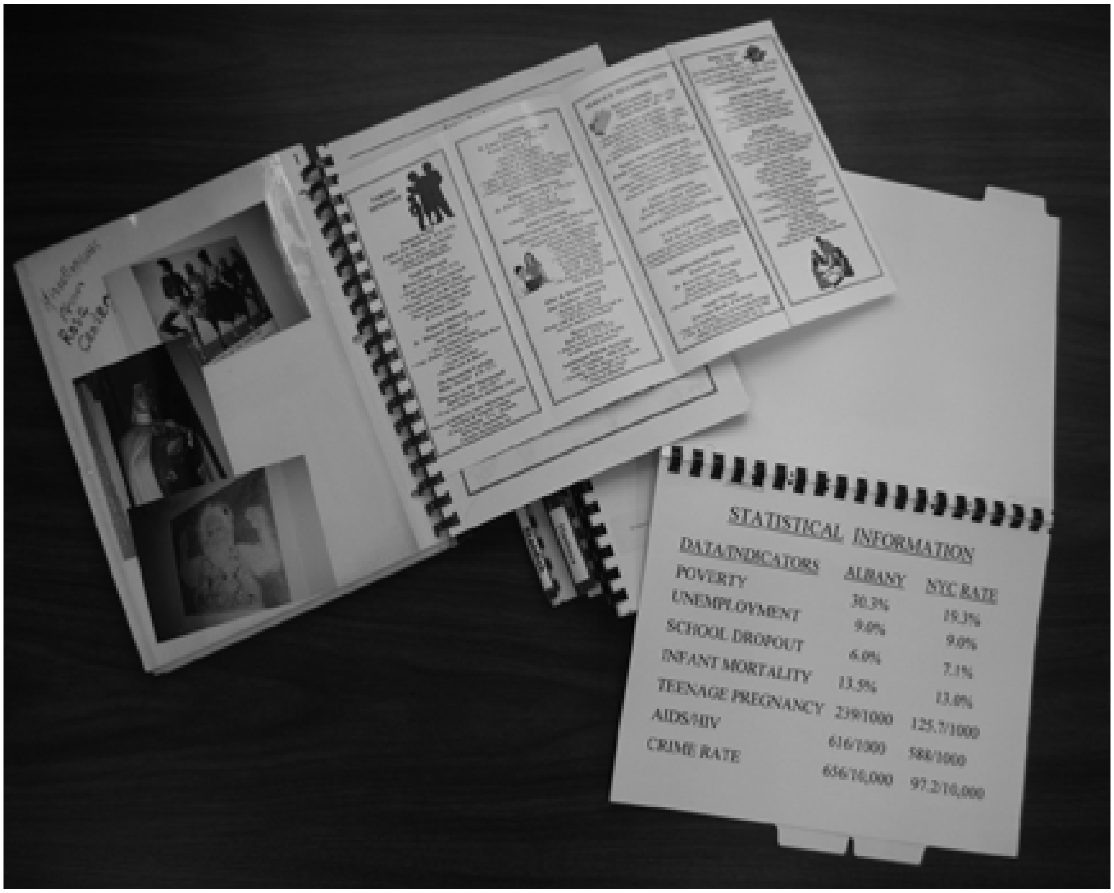
Table 1

Dublin Core Qualified Element	Description
dc.contributor.advisor	Faculty advisor associated with theses/dissertations/student projects.
dc.contributor.author	Author of the resource.
dc.contributor.committeeMember	Faculty committee member associated with theses/dissertations/student projects.
dc.contributor.editor	Editor of the resource.
dc.contributor	Catch-all for unspecified contributors.
dc.date.accessioned	Automated by Dspace. Date Dspace takes possession of item.
dc.date.available	Automated by Dspace. Date item becomes available to the public.
dc.date.copyright	Date of copyright if different from date.issued.
dc.date.issued	Date of formal issuance (e.g., publication) of the resource.
dc.identifier.bibliographicCitation	Standard bibliographic citation (APA style).
dc.description.abstract	Abstract or summary.
description.degree	Indicate degree awarded; for example, Master of Arts, Master of Business Administration, Master of Fine Arts, Doctor of Business Administration.
description.school	Academic School associated with the resource.
description.program	Academic Program associated with the resource.
dc.description.tableOfContents	Table of Contents of the resource.
dc.description.provenance	Automated by DSpace. History of the custody of the item since its creation.
dc.description	Catch-all for a general description of the resource.
dc.digSpecs	Technical information about the hardware, software, and processes used to create the digitized resource.
dc.format.extent	The size or duration of the resource. (When describing digital resources, best practices are to use file size in bytes).
dc.format.mediaType	A file format or physical medium. (example: PDF/A)
dc.identifier.isbn	ISBN
dc.identifier.issn	ISSN
dc.identifier.uri	Automated by DSpace. The Uniform Resource Identifier.

dc.language.iso	Language of the resource.
dc.publisher	Entity responsible for publication, distribution, or imprint.
dc.references	A related resource that is referenced, cited, or otherwise pointed to by the described resource.
dc.relation.hasPart	A related resource that is included either physically or logically in the described resource.
dc.relation.hasVersion	A related resource that is a version, edition, or adaptation of the described resource.
dc.relation.isPartOf	A related resource in which the described resource is physically or logically included.
dc.relation.requires	Referenced resource required to support delivery or function of item. (For example, Acrobat Reader.)
dc.rightsHolder	The owner of the copyright.
dc.source	A related resource from which the described resource is derived.
dc.subject.other	Local controlled vocabulary; for example, keywords. We should always include the applicable geographic name(s).
dc.subject.lcsh	Library of Congress Subject Headings.
dc.title	Title of the resource.
dc.title.alternative	Varying (or substitute) form of title proper appearing in item, e.g. abbreviation or translation.
dc.type	Nature of genre of content. Use MARC Value List for Genre Terms; for example, "thesis".

Dublin Core element definitions are generally according to the DCMI Metadata Terms, with local enhancements made for our repository's particular needs. [6]

Figure 1



STATISTICAL INFORMATION

DATA/INDICATORS	ALBANY	NYC RATE
POVERTY	30.3%	19.3%
UNEMPLOYMENT	9.0%	9.0%
SCHOOL DROPOUT	6.0%	7.1%
INFANT MORTALITY	13.5%	13.0%
TEENAGE PREGNANCY	239/1000	125.7/1000
AIDS/HIV	616/1000	588/1000
CRIME RATE	656/10,000	97.2/10,000

Examples of the various materials included in the SCED student thesis projects.

Endnotes

[1] For more information about DSpace, visit http://www.dspace.org/.

[2] Edward Mendelson, "ABBYY FineReader Professional 9.0," *PC Magazine* (April 18, 2008), http://www.pcmag.com/ article2/0,2817,2305621,00.asp, and Edward Mendelson, "OmniPage Professional 16," *PC Magazine* (April 15, 2008), http://www.pcmag.com/article2/0,2817,2305588,00.asp.

[3] Julie Allinson, "Describing Scholarly Works with Dublin Core: A Functional Approach," *Library Trends* 57, no. 2 (Fall 2008): 221-243.

[4] CDP Digital Imaging Best Practices Working Group, "BCR's CDP Digital Imaging Best Practices Version 2.0," *BCR* (June 2008), http://www.bcr.org/dps/cdp/best/digitalimaging bp.pdf.

[5] Library of Congress, "PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4," *National Digital Information Infrastructure & Preservation Program* (March 7, 2007), http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml.

[6] DCMI Usage Board, “DCMI Metadata Terms,” *DCMI* (January 14, 2008), http://dublincore.org/documents/dcmi-terms/.