Digital Retrospective Conversion of Theses and Dissertations

ETD2005 Conference

Digital Retrospective Conversion of Theses and Dissertations: an In House Project

Todd M. Mundle, Associate University Librarian
Simon Fraser University
Burnaby, BC Canada

Todd M. Mundle
Associate University Librarian, Simon Fraser University Library,
Burnaby, BC Canada

Keywords: Digitization, retrospective, in-house, Canada

ABSTRACT

Realizing the importance of easy access to and the promotion of the scholarly output of Simon Fraser University (Burnaby, BC, Canada) graduate students, the SFU Library began exploring a retrospective digitization plan for theses and dissertations. Starting in 2004 and knowing the cost of an outside agency to complete the work, the SFU Library explored establishing an in house retrospective plan to digitize the approximate 4500 theses and dissertations produced between 1965 and 1996.

With the blessing of the Dean of Graduate Studies a cost analysis per title was completed, special project money secured and a coordinator hired to work on this and other digitization projects for the Library. Using student assistants, work commenced in the fall of 2004 with existing microfilm and paper theses. The items are digitized in PDF and stored within the Library's institutional repository using DSpace. Metadata is created from existing library catalogue records and the titles will also be linked from the catalogue. The project has shown to be cost effective when compared to the title by title costs provided by an outside agency.

A natural progression from this project was creating PDF files of current theses and dissertations from the unbound paper copies. Since December 2004 these titles have been added to the workflow including metadata created from a spreadsheet of information maintained by the Thesis Assistant. For current titles, copyright licences are signed by students when submitting final copies. The Library and office of the Dean of Graduate Studies are in the process of securing agreement for the retrospective titles.

This paper explores the processes used to establish the project and discuss how it could be implemented at other institutions in a cost effective manner.

Digital Retrospective Conversion of Theses and Dissertations

1. INTRODUCTION

Simon Fraser University (SFU) is a medium sized comprehensive university located on the west coast of Canada in the province of British Columbia. It was started in 1965 and offers undergraduate and graduate programs at the master's and PhD level. Since 1967 students have been producing theses and dissertations of which approximately 4500 were completed between 1967 and 1997. Retrospective digitization of these 4500 is the focus of this paper.

At SFU, the physical handling and receipt of theses is done by the Library. The SFU Library has a number of years experience in creating digital versions of special collections and material for which copyright had expired. Part of its digitization plans included investigating digitizing both retrospective and new theses. So when approached by UMI Dissertation Publishing (UMI) in the fall of 2003 with a proposal to digitize the back collection of theses and dissertation, it took a much closer look at what would be involved by doing it in house.

The UMI offer was alluring in that it offered a solution that was clean and quick. Using microfiche copies from UMI's vault and scanning of any paper only titles, UMI proposed to undertake this conversion process and deliver to SFU digital files and free MARC records. A number of flags were raised when we investigated further: 1. Although SFU would be provided with digital copies, UMI would retain a copy on their servers in the United States; 2. Access to the digital copies would be restricted to current SFU faculty, students and staff only; 3. SFU is not the copyright holder so can we legally digitize the titles?; and 4. For the price we would pay UMI, could we do it using local resources? While we investigated these issues, we told UMI that the proposal was on hold.

We did an initial cost analysis using data from other digitization projects involving microfiche digitization. As SFU already owned a high end microfiche scanner the bulk of the costs would come from labour and administration of the project. Early indication was that it would be slightly less expensive for us to do this work locally but that the benefits would offset the investment: 1. local control over digitization; 2. hiring SFU students to do the work thereby keeping the money local; 3. wider access to SFU scholarship by offering the digitized copies beyond SFU.

Before committing dollars to the project we needed to investigate the legal aspects. How does copyright impact on our ability to digitize? How would we control copying and editing of the works? How do we cover off our responsibility to the copyright holders (the students) while offering their scholarship to a wider audience? We spoke with SFU's legal counsel and were advised to do the following:

1. The SFU Library should set up a Thesis Access Policy stating that the theses will be made available for research purposes or private study. Such a policy will bring the issue within the fair dealing exemption in s.29.1 of the Canadian Copyright Act. The Library should then post an appropriate notice of this policy on the website.
2. The SFU Library should contact theses authors where possible giving them the opting out option.
3. When digitizing, use PDF security measures to prohibit copying/editing/printing.

The Thesis Access Policy has been established and we are in the process of contacting the authors. Since it will take some time to contact the authors we decided to start the digitization and wait to release the digitized versions until either after a period of time or when contacted by the author. As such we began to digitize in the fall of 2004.

Digital Retrospective Conversion of Theses and Dissertations

2. METHODS

Scanning is done from the microfiche copy when available or from paper if either the microfiche quality is not good enough or is not available. Students perform an initial physical scan of the fiche copy to review that the quality is appropriate and to check to see if an abundance of pictures, graphs, etc. exists that may not scan well. Practice has shown that the quality is consistent throughout a single piece of fiche therefore only the first few pages need to be checked. If the fiche fails this step a digital copy is made from the paper thesis. However if the fiche passes it is scanned using a Canon microfilm scanner in a batch format. Each page of the thesis is scanned as multiple BMP file and stored in one folder for each thesis. To maintain control the folders are named using .bnumbers from the library's Innovative Interfaces (III) catalogue. The .bnumber ties the digital file to the library catalogue MARC record and is unique.

The folders of BMP files are then converted to TIFF files. These files are then checked for quality and processed involving: erasure of signatures on approval pages to comply with Canadian privacy laws; cropping of pages using PhotoShop; re-scanning of pages to improve the quality. The files are then converted to PDF using Adobe Acrobat. To make them more useful, the PDF pages are then "captured" using Adobe Acrobat to make them keyword searchable. The students then create the metadata for each thesis by using cut and paste taking author, title and subjects from the library catalogue record. The PDF security measures are then put in place allowing users to view but not print or edit the documents. The final step at this stage is a quality control check confirming all pages are scanned, random quality check of pages, testing keyword search capability and insertion of grey scale images if necessary. If the scanning is required from the print version, the various steps are the same except that using the flatbed scanner, TIFF files can be created from the original scan bypassing the need to convert BMP files to TIFF. Otherwise the processing, capturing, security control and quality assurance steps are identical to the fiche scanning process. The documents are now ready to be uploaded into the SFU Library's Institutional Repository.

The SFU Library is using DSpace for its institutional repository. Since the retrospective theses all have MARC records in the library catalogue these records are used as the metadata for the DSpace record for each thesis. Using a Perl script (marc2dspace.pl) the MARC records are imported into DSpace and attached to each PDF file. The files are linked together by the unique .bnumber from the MARC record. Using the DSpace import utility, the PDF file with metadata is imported into DSpace and put into the "Thesis, Dissertations and other Required Graduate Degree Essays" community. The items are searchable by single keyword from the title, author, and abstract. As well the community is browsable by title, author or date. Once the PDF is opened it is then keyword searchable using Acrobat.

The next step is to put the URL for the DSpace thesis back into the MARC record in the library catalogue. Again using the DSpace import utility a DSpace map file is created taking the .bnumber and handles or locations from the DSpace record. Using another Perl script (updatethesesmarc.pl) brief MARC records consisting only of 035 (for .bnumber) and 856 (for URL) are created. These records can be overlaid onto the existing records in III since .bnumber overlays are reliable. Once set up these scripts can be run without human involvement.

An added bonus from doing this project was that as of December 2004 we have begun scanning PDF versions of the current theses before they get sent off to Library and Archives Canada and ultimately to UMI for microfiche and PDF production. The unbound theses are document fed into a flatbed scanner that can scan a 150 page thesis in 8 minutes. The processing then takes the same path as the retrospective theses except for the creation of the DSpace record and the record to be loaded into the library catalogue.

Records for DSpace are created from an excel spreadsheet produced by the Thesis Office and then linked to the PDF using a unique "etd ID" tag because the .bnumber does not exist. Then to create a brief MARC record for the library catalogue, another Perl script (dspace2marc.pl) is run which takes the spreadsheet information and the DSpace map file to create a record containing the DSpace handles (URL), author, title, department, and abstract. These records are then loaded into the catalogue where subject headings are added and authority control is done.

To cover off the copyright concerns, for the new theses, students sign a partial copyright license allowing the SFU Library to create digitized versions of their work for disseminating through the SFU Library Institutional Repository.

3. RESULTS

As of June 2005 we have digitized 795 retrospective theses, and 650 new theses including many that do not go through the LAC/UMI process. On average it costs $19.40 CDN ($20.36 AUS; $15.49 US) to digitize each thesis. This is approximately 25% cheaper than had we gone with the original UMI proposal.

It may take us longer than it would UMI to complete the project but we expect to be done by the end of 2006. Time and cost have been affected by two issues:

1. High incidence of photos, detailed diagrams, etc. in the body of the theses, which slows down scanning. This may become less of an issue as we go back further in time.

2. Some incidence of large fold-put maps that required considerable time to scan. We may consider not scanning this type of appendix in the future.

4. CONCLUSION

This project has provided us with three positive results:

1. A process to digitize scholarship that before had only a limited audience and distribution;
2. Spending money locally to hire students to do the digitization; and.
3. Ongoing commitment to digitize “new” theses.

Without the impetus and experience of doing the retrospective work, we would not have developed a plan for future theses. That will be the legacy of this project. That and the fact that we took the path that was best for both our current students by hiring them to do the digitization and alumni for providing a wider distribution mechanism for their scholarship.