Dissertations Unbound: A Case Study for Revitalizing Access

Gary M. Worley, Ph.D.

Digital Imaging, Information Technology, Virginia Polytechnic Institute and State University

Abstract

Universities, even those with active electronic theses and dissertation projects, often ignore a larger set of research documents housed within their walls, the printed and bound dissertations usually found within the campus library. For most universities, these bound dissertations constitute the majority of the documents associated with graduate student research. Access to graduate research documents is then divided between the online electronic dissertations and the printed dissertations stored neatly in rows on numerous shelves. This separation for access represents a dichotomy of scale, with an electronic dissertation offering unlimited access by many users at the same time versus the bound dissertations with limited access available to one user at a time. A simple solution is to have the bound dissertations digitized. But, what does this really mean? And, how long does it take at what cost? A primary concern for undertaking such a scanning project centers on the quality of the product versus the cost associated with the production of the digital files. One consideration is to simply outsource the project. This approach, however, may result in a compromise of the quality for increased speed in delivery. An alternative approach is campus-based where support is provided and long-term storage/preservation is addressed. This paper explores the various aspects involved with the support of a large-scale dissertation-scanning project, and the production process adopted at Virginia Tech for the development and storage of the scanned dissertations targeted for inclusion in the ETD database on our campus.

INTRODUCTION

Beginning with the 1996 academic year, graduate students at Virginia Polytechnic Institute and State University (Virginia Tech) were required to submit their graduate theses and dissertations to the graduate school in electronic format. This requirement represents a significant change for the retention of these manuscripts at the university by removing the need to store graduate research documents on the countless shelves they would have traditionally occupied to computerized storage devices capable of delivering those documents through a networked database. As a result, the electronic theses and dissertations (ETDs) represent a resource for research data and information that can easily be shared by many users simultaneously. Restrictions for access are imposed only at the author's request and as each document is released, the ETD becomes available to a world audience.

In contrast to the more than 9,000 electronic theses and dissertations, the University Library also maintains over 30,000 printed volumes representing the work of students who were required to submit a printed and bound manuscript describing their graduate research. These documents share the same issues as their counterparts on shelves at other universities, with limits for access and restrictions associated with their removal to off-campus sites. In many instances these documents are never touched once they assume their place on a shelf, and if used, their availability is consumed wholly by singular access to the person holding the document in hand.

Limited access may also influence the level of use for these types of documents. This influence was indicated in 1999 as part of an evaluation comparing the circulation of the bound theses and dissertations to the newly implemented ETDs at Virginia Tech. Based on the combined average circulation for both groups, the bound theses and dissertations were determined to be relatively inaccessible and under utilized. For example, theses submitted between 1990 and 1994 had a combined average circulation of 2.24 per year for each document, dissertations for the same period were only slightly higher at 3.2 per year for each copy. In contrast, the combined circulation for ETDs from 1996-1999 was 270 for each document per year (McMillan, 2001). This comparison represents a significant departure from the traditional access to these types of documents in favor of the digital versions. In addition to these findings, a more closely controlled

circulation test was conducted between 2005 and 2006 comparing a select group of dissertations that were available as bound volumes in 2005 and as digital PDF files in 2006. The average circulation for these documents differed greatly from one year to the other, with the average combined circulation in 2005 at 11.5 per document and the average circulation in 2006 increasing to 126.6 per document. This observed increase in the combined circulation should not be interpreted as an indication that all the documents are affected equally (Table 1.), however, it does provide evidence that by altering the method of access for the bound theses and dissertations, there is definite potential for increasing overall circulation.

Table 1. Circulation comparisons for 10 dissertations between 2005 and 2006.

	YEAR		AUTHOR NAME		
URN	2006*	2005**	LAST	FIRST	TITLE
02032004-161558	275	11	Jeffrey	Thomas J.	Adaptation and validation of a technology attitude scale for use by American teachers at the middle school level
02032004-161557	0	6	Bonfadini	John	Assessing and Changing the student teacher and his learning environment with student ratings and peer group counseling sessions
02032004-161657	118	17	Herbert	George Robert	Comparative analysis of personality characteristics of industrial arts teachers in the United States.
02032004-161628	319	4	Mack	Warren	Effect of Training on Computer-Aided Design on the Spatial Visualization Ability in Selected Gifted Adolescents
02032004-161548	0	10	Holmes	James A.	Formatting variables and typeface variations of dot-matrix print and their effect on reading comprehension and reading speed
02032004-161618	0	3	Lee	Bruce Tien-Lung	Meta-evaluation of Taiwan Ministry of Education's National Technology Institutes evaluation: A Study of Evaluation Team's and Stakeholders' Judgments on the evaluation practice
02032004-161617	134	54	Copeland	Leon L.	National Study using the Delphi technique to identify teacher competencies for evaluating industrial arts student teachers
02032004-161638	0	1	McCade	Joseph M.	Projecting acceptance into Millersville University's Department of Industry and Technology Using High School Rank, Social Capital, SAT Scores, Sex, Age, and Race
02032004-161648	210	4	Meide	Jeff	Pupil's attitude toward Technology Bostswana
02032004-161607	210	5	Cooper	Joseph Linwood	Study of the Effects of cognitive training on the ability of adolescent educable mentally retarded students to learn and retain vocational competencies
Average	126.60	11.50			

* fy06 web hits on ETD only (i.e., pdf)

** fy05 Addison checkouts since 6/19/5

As demand for access to information online increases, the limitations associated with printed documents have become more pronounced. This is especially true for theses and dissertations with very few copies in existence to make available for distribution (Eaton, 2001). Given these limitations and our observations as described above, it was not difficult to reach a consensus on our campus to migrate the printed documents to digital formats suitable for inclusion in the existing ETD database. The problem then becomes one of defining the process for preparing the materials and creating a management workflow for eventual access to the digital files, all while imposing the least impact on the staff and users associated with those materials.

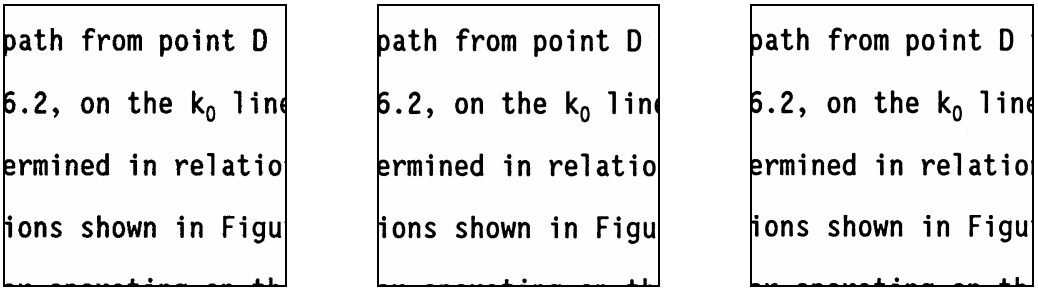
FILE SPECIFICATIONS

The first consideration with any scanning project focuses on the intended use for the digitized product. This dictates the target for the resolution of the scans and should make allowances for any intended post processing of the files. ETDs on our campus are submitted as Portable Document Files (PDF) and often include other media file types referenced within the PDF document. The printed dissertations do not include multiple media formats other than printed pages, photographs, and drawings. In itself this limitation simplifies projects of this type by narrowing the resolution issues to images and text. With PDF established as the delivery format for the digitized documents, the next step is defining the required resolution for our target, or intended use.

A common oversight associated with materials selected for scanning concerns the resolution for the digital file. Scanners typically provide capture capabilities within a range of settings for bit-depth and number of pixels per linear measure. These settings are configured to accommodate the requirements for viewing and for reproducing the scanned materials. The initial scan should then include allowances for the greatest resolution need anticipated; however, resolution needs should not be interpreted globally as a reason to configure a scanner to the highest resolution settings possible for the device. Scans created at higher resolution settings require more storage space and depending on the original material do not necessarily result in better image captures.

For the purposes of this project the digitized versions of the drawings and photographs included in the documents are handled separately from the pages containing only text or line art. This is due to the resolution demands specific to the continuous tone images, which are quite different than that for black text and line art. Typically grayscale, or images represented by a range of tonal value, require 8-bits or 256 values of black-and-white variation. With 8-bits of rendering depth, these types of images can be scanned at 300 ppi (pixels per inch) to accommodate a wide range of media requirements for display and printing in order to render the entire image or portions of the image representing detailed information. For this reason, images of photographs and drawings included as part of the printed thesis and dissertations are scanned separately and reinserted into the PDF file appropriately within the text pages where they appeared in the printed documents. For the black-and-white page elements, such as text and line drawings, the optimal scanning bit depth is 1-bit. With the absence of grayscale values, 8-bits of tonal rendering are not necessary for the text or for the line art. The number of pixels describing the black-and-white areas such as the text, however, is important and should be sufficient to meet the needs of the highest post-processing target (Figure 1.).

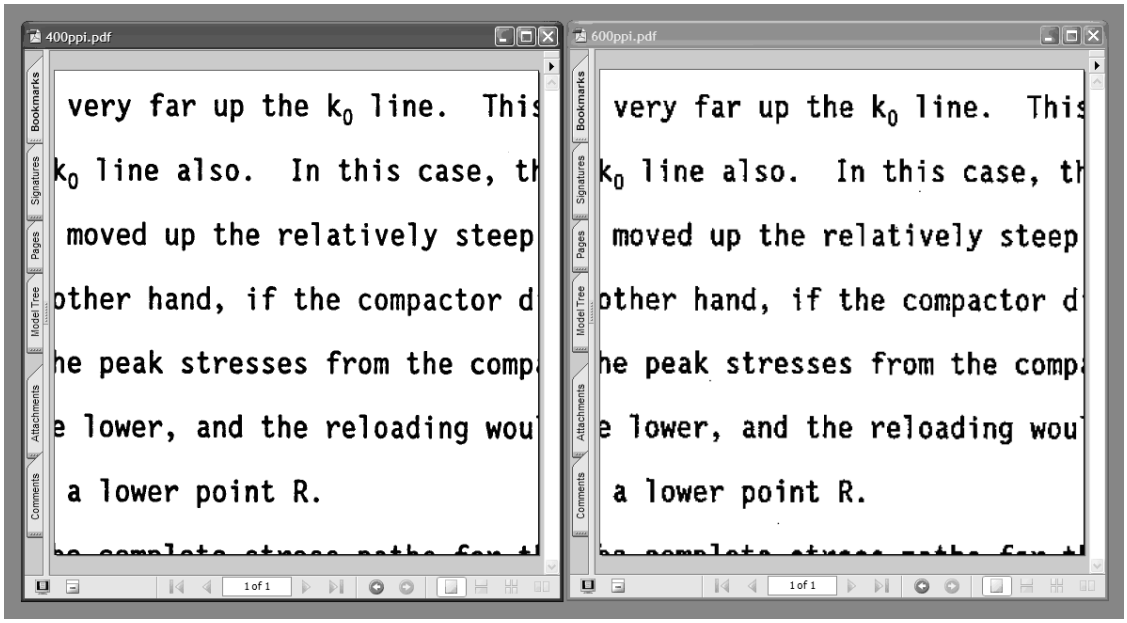
Figure 1. Scanning comparison for a black-and-white text image with three different values for the pixels per inch (ppi). The bit depth is constant for the three samples.



300 ppi, 52 Kb 600 ppi, 126 Kb 1200 ppi, 422 Kb

The scanned samples presented in Figure 1 indicate no visible difference for the text characters at 1200 ppi and 600 ppi. There is a slight decrease in quality at 300 ppi indicating that a usable resolution for meeting the needs of the project exists between 300 ppi and 600 ppi. The following examples (Figure 2.) both demonstrate adequate image definition for viewing the document text and the resulting PDF files, and with JPEG compression applied are also similar in size, at approximately 70 Kb. Given the similar result for both 400 ppi scans and 600 ppi scans, the lower resolution setting can be used for scanning the text, which also works within the technical specifications for the scanners available to the project, a Fujitsu fi-4750c document scanner for the grayscale pages, and a Fujitsu M4099D document scanner for the black-and-white pages at 400 x 400 pixels.

Figure 2. Resolution comparison for a black-and-white text image scanned at 400 ppi and 600 ppi with JPEG compression applied to create PDF samples.



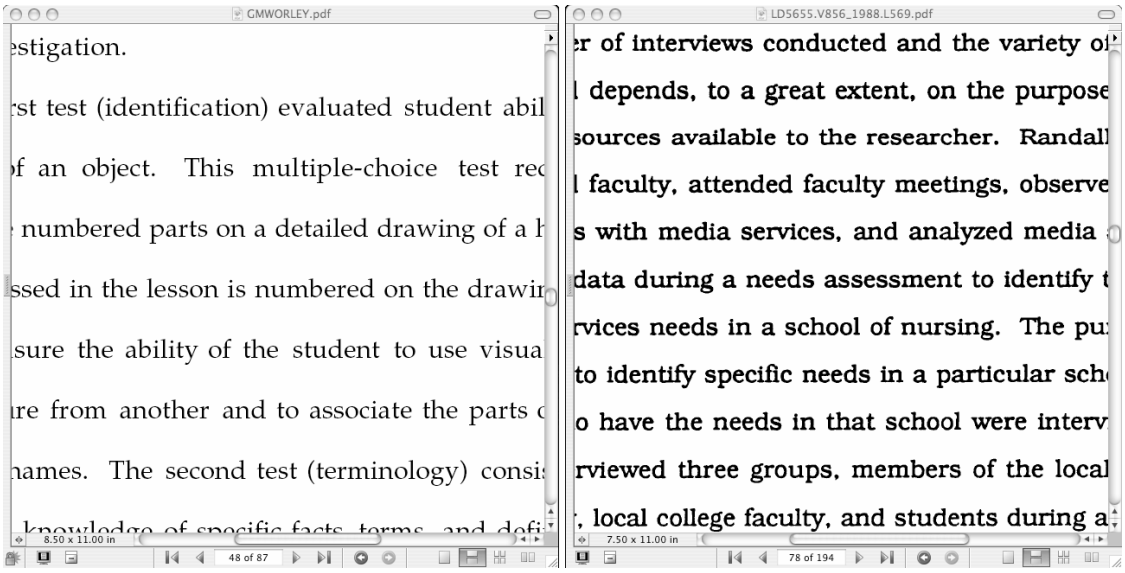
Scanned at 400 ppi Scanned at 600 ppi

An important requirement for creating digital versions of the printed theses and dissertations for this project is to provide a capability for possible optical character recognition (OCR), seen as a functional need for text conversion. The minimum resolution required now with PDF (Acrobat 7.0) for applying OCR is 144 dpi and this includes the applied compression when the file is created. For the purposes of this project, the OCR requirement became the target for the resolution settings and can be achieved by applying JPEG compression set to 150 dpi when creating the PDF files.

The choice to OCR the text as part of the production process was not a consideration for our project. The type in the scanned theses and dissertations can be converted to text characters, however this introduces the need to proofread each page, adding significantly to the production process. An alternative to this is allowing the user to perform the OCR on the document when needed as the document is used. This capability is available to users either through Microsoft Office Document Imaging software, or through Adobe Acrobat Professional. For either application, the user can verify the accuracy of the applied OCR by using the original document image as a reference. Any portion of the document can be processed in this way, or the entire document can be processed if text searching is required. In either case the OCR accuracy is verified by the user, thus eliminating a significant cost to the project for this function.

Comparisons between a sample ETD and a sample bound dissertation that was scanned indicate no significant differences with text legibility for the two types of documents (Figure 3.). As shown in the following examples, the scanned text appears on screen in a similar fashion to the electronic text. This similarity also applies if the two documents are printed on a laser printer. The major difference for the text in the two documents is that the electronic text can be selected within the document making it possible to search the text or make selections for copying and pasting into another application without the need to convert the text using an OCR application.

Figure 3. Display of black-and-white text comparing a typed PDF file with a PDF file created from a scanned page.



Electronic Text Scanned Text

Based on the average file size for each digitized page, approximately 70 Kb, the minimum storage requirements for the project were estimated at 13 Mb per document. The document size was estimated using 186 pages to represent the average for the bound theses/dissertations, including the appendices. Given the approximate number of theses and dissertations, 30,000, the storage requirements for the completed project then are estimated at approximately 390,000 Mb, or 390 Gb of storage space. Based on this projection, a conservative estimate for planning the available storage would be 500 Gb, or .5 Tb of disk space.

With the current low cost for storage media, the estimate for the total project does not represent a significant investment and can be easily accommodated within the current ETD database capacity. Backups are provided through the Meta Archive Preservation Initiative, which provides participating institutions the ability to store multiple copies at multiple locations. This initiative is a multi-institution approach for providing archival backup and storage of digital files so that copies are maintained off-site

from the primary source locations. An additional backup is also maintained in Digital Imaging as part of the Library Archives housed on their departmental server.

PRODUCTION WORKFLOW

Selecting a starting point for where to begin this project, while still accommodating requests to checkout the bound theses and dissertations, presents several different possibilities. In order to maintain the linear quality of the existing ETD database, we first considered working in consecutive order on a year-by-year basis either by moving from the oldest documents to the most recent, or from the most recent to the oldest. Starting with the oldest documents offered the opportunity to gain some experience for the project, and make adjustments to the workflow prior to handling those volumes that might be in demand more frequently. Starting with the most recent documents, however, offered the opportunity to backfill the ETD database and as a result, increase the consecutive years represented in that online collection. Another option with less appeal was to group the documents by academic area and complete each group based on demand for that area of study. In the end, none of these seemed to satisfy our immediate goal to make the printed documents in demand more accessible to multiple users on our campus. In order to address this concern and as a way to begin the process with minimal disruption, we finally decided to approach the documents already in use as they were returned to the library through the Circulation Department. In part this was based on the fact that as these documents returned to the library they could be scanned and processed within the normal timeframe allocated for their return to the shelves, but more importantly, the documents in circulation demonstrated a higher demand for current use.

As a result, the project follows a simple course for selecting the thesis or dissertation to scan. If the thesis/dissertation documents have been checked out, they are grouped as part of the project when they return to the library circulation desk. On a weekly basis, those documents that have been returned are grouped into sets of fifty. If the weekly total for the grouped sets is less than fifty, additional thesis or dissertations are then selected from the shelved volumes on a systematic basis moving backward in time by date. By taking this approach we were able to establish a minimum number that each department could use for scheduling the workflow in their respective areas.

With the method for document selection in place, we next identified the different aspects of the workflow and the departments involved with processing the materials. This workflow involves 3 departments within University Libraries and two additional departments in Information Technology. The departments in University Libraries are responsible for tracking the theses/dissertations as the printed versions return to the library or are removed from the shelves. This includes recording the bibliographic records for the printed documents and making the appropriate changes to that record as the digital files are made available within the ETD database. Library staff also select and organize the materials for processing including pickup and delivery of the processed documents. The two departments in Information Technology perform separate functions: University Printing Services, responsible for separating and trimming the pages for each thesis and dissertation; and Digital Imaging, responsible for scanning the material and creating the PDF files from the documents, providing high quality production resources in support of the project.

Beginning with the **Circulation Department** in University Libraries, the printed documents are processed as follows:

Step 01) The thesis/dissertation is checked in as returned, or removed from the shelves

Step 02) The Barcode is deleted, and the bibliographic ID is noted

Step 03) The covers are removed, and the documents are identified and grouped in sets (50)

Step 04) The bibliographic ID list is sent to Library Systems

Step 05) The Document sets are delivered to Printing Services for trimming

Upon receiving the bibliographic ID, **Library Systems** extracts that information from Addison (Appendix 1.) and awaits further notification that the volume has been scanned and the PDF file is copied to the ETD database.

Step 06) The Bibliographic IDs are extracted from Addison

The document sets are then delivered to **University Printing Services** where the binding is trimmed away creating separate pages for each volume. The pages for each document are trimmed to 7.5 inches, which works within the margins used for these types of manuscripts and the page requirements for the document scanners available to the project.

Step 07) Document binding is trimmed away from the pages

The **Circulation Department** makes arrangements with Printing Services to schedule the pick-up and delivery of the document sets that require trimming. As this occurs the trimmed documents are then transferred to Digital Imaging for scanning.

Step 08) The trimmed volumes are picked up from Printing Services

Step 09) And delivered to Digital Imaging

The **Digital Imaging Center** provides a wide range of scanning services to the University including document scanning and flat art scanning. These services are provided in support of approved projects at no charge.

Step 10) The document sets are reviewed for the types of materials included

Step 11) Each document page is then scanned for text

Step 12) Image pages are then re-scanned as grayscale and inserted as replacement pages

Step 13) The completed document is assembled into a PDF file, and a quality check is performed

Step 14) As each document set is completed, Digital Library and Archives is notified

The **Digital Library and Archives** (DLA) maintains systems including those for electronic theses and dissertations and digital images. With the completion of scanning and PDF file development, each document set is returned to DLA and the electronic (PDF) files are transferred.

Step 15) The theses and dissertations are transferred from Digital Imaging

Step 16) PDF files are reviewed for completeness

Step 17) As necessary, notifies Digital Imaging for corrections

Step 18) Notifies Circulation of completed set reviews

Step 19) Copies PDF files to ETD database

Step 20) Notifies Library Systems

After receiving notification that the PDF file is copied to the ETD database, **Library Systems** updates the MARC record and establishes the HTML link to the ETD database.

Step 21) Grabs MARC

Step 22) Turns MARC into HTML record for ETD database

Step 23) Links HTML with PDF

Step 24) Notifies Cataloging matching MARC record to URL

As notification is received that an HTML link is established for the PDF file, the Circulation Department updates the MARC record (Appendix 1.) with the URL making the transfer from a bound document to the ETD database complete.

Step 25) Updates appropriate MARC record with URL and systems notes

RESULTS AND CONCLUSIONS

With over 30,000 printed theses and dissertations to process, time becomes a significant factor for judging progress with this project. Averaging 100 volumes per week, the entire collection of bound theses and dissertations could be completed within 6 years. An average of 50 per week would extend the project timeline to 12 years. Based on these estimates, the approach adopted for the project assumed a slow start for the number of volumes we would average per week. However, we also assumed that as the project moved forward and our experience increased, we could increase the production to a sufficient level so that the completion date for the project would fall somewhere between the 6 year and 12 year estimates. To date over 1,500 theses and dissertations have been digitized representing 258,563 pages. This indicates an average of 20 volumes per week since the project start date of July 2005, and represents a number well below the minimum required to complete the project in less than 12 years.

The benefits for undertaking such a project are immense. Providing the research online has allowed the university to increase access for this information to students and faculty at the university. It also provides direct support to the land-grant mission for the university through increased opportunities to share information. Even so, attempts to maintain a steady workflow have encountered myriad problems in terms of transferring the materials and coordinating schedules for the work. Initially, we believed that the volumes returning through the Circulation Department each week would provide a sufficient core of documents to sustain the minimum processing levels required for the project. To supplement that number, additional volumes would be selected from the remaining shelved thesis and dissertations. As these documents were selected, they would begin to move through the processing steps as outlined previously.

Staffing limitations, diverse schedules, and other priorities have all contributed in one way or another to limit the number of theses and dissertations moving into the workflow. This results in delays for the other departments contributing to the project and often results in a backlog for one area or no work at all somewhere else. For example, the Printing Services department, responsible for trimming the bindings and creating the separated pages, requires a minimum of 100 documents to justify scheduling that activity. If the Circulation Department cannot prepare a sufficient number of volumes to transfer to Printing, then the trimming activity is delayed until a sufficient number does arrive. These types of delays result in a gap for the delivery of documents to the Digital Imaging Department and further delay the transfer of the scanned digital files to the Digital Library and Archives. This cascading effect creates an inconsistent workflow and this impacts scheduling for every department.

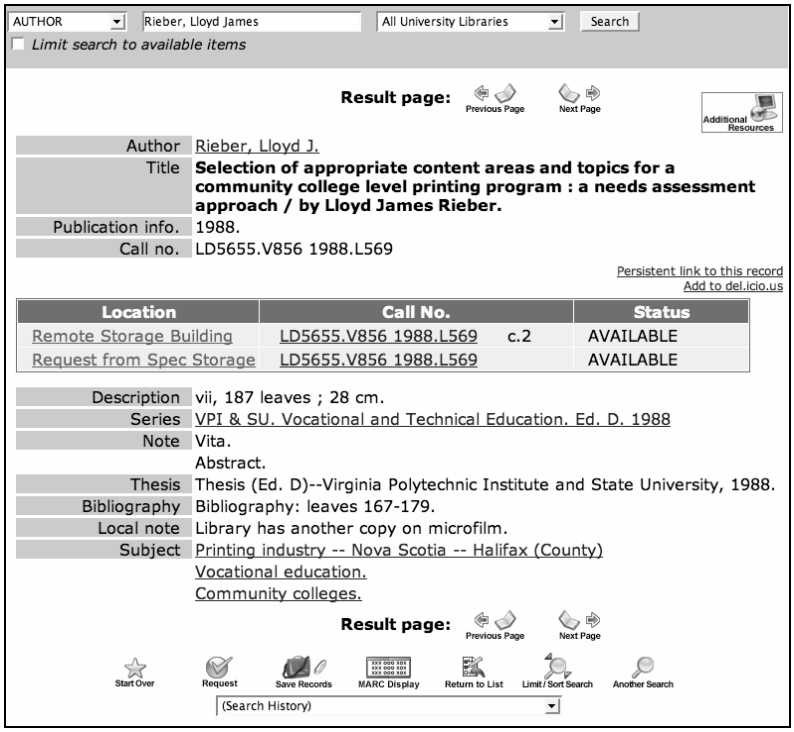
One obvious lesson we learned is that no matter what production process is adopted, the scanning function can easily overwhelm the capacity for all the other units involved with the process. It is quite possible that all 30,000 volumes could be scanned in a short period of time, but that in no way addresses the steps required to adjust the MARC records, or the issues related to organizing the digital files for transfer into the ETD database. Many of the thesis/dissertation documents include unique representations in the form of drawings or diagrams and these most often are included as loose papers stored in a pocket attached to the cover. The accompanying materials also represent oversized paper dimensions requiring the use of scanners other than the document scanners used for the letter-sized pages. This feature alone calls attention to the need to treat the documents in a careful manner and is one argument against a speedy approach.

Other factors that impact scanning involve unexpected elements. For example, dissertations that include photographs glued to a sheet of paper, or examples of materials such as a piece of cloth, all require additional staff time for scanning individually. Another example is the use of confidential information that the author may have included as part of the title page of the dissertation. At the time many of the bound dissertations were submitted, the inclusion of personal information was not considered a threat to the author's identity. Sitting on a shelf, these dissertations were also protected to some extent because of the limitations for access. Unfortunately, as a digital document that security is no longer the case. This discovery slowed our process dramatically while awaiting a decision by the Graduate School for approval to alter the title pages accordingly by removing any personal information. Although the approval was eventually made, an additional step was introduced to review the documents for personal data, and this again affected our progress.

The promise and potential for the digitized bound theses and dissertations remains the driving force for this highly regarded effort on our campus. As awareness for the project has increased, we are also beginning to receive requests for specific groups of topics to be included. This is currently being considered as an additional source for improving the selection process, viewed as the single most important element in terms of maintaining a consistent workflow for the project.

As these issues are resolved, the project overall is experiencing a slow but steady increase for the numbers of documents processed each week. Continued progress will eventually result in an ETD database that represents the entire set of theses and dissertations produced at Virginia Tech, providing a rich resource to the campus community and a world audience.

Appendix 1. MARC record for Addison reference before and after the theses/dissertation is added to the ETD database.



REFERENCES

Avedon, D. (1997). Quality Control of Electronic Images. Silver Spring, MD: Association for Information and Image Management International.

Eaton, J. (2001, December). Why digital theses and dissertations? How can you get started. Presentation at the Council of Graduate Schools National Meeting, San Diego, CA.

Kleper, M.L. (1987). The Illustrated Handbook of Desktop Publishing and Typesetting. Pittsford, NY: Graphic Dimensions.

McMillan, G.M. (2001, October). Access to EResources: theses come out of the attic. Presentation at the meeting of the Virginia Library Association, Richmond, VA.

Tufte, E.R. (1990). Envisioning Information. Cheshire, CT: Graphics Press.