

Dissertation Archiving and Access

*A case study for
accessibility and preservation*

A report from:

ProQuest[®]
Start here.

Dissertation Archiving and Access: A Case Study for Accessibility and Preservation

Austin J. McLean
ProQuest
Ann Arbor, MI/USA

ProQuest
789 E. Eisenhower Parkway, Ann Arbor, MI 48106-1346
800-521-0600

Abstract

Many universities keep paper copies of dissertations, without reliable back-up. Vulnerable to theft, fire and decay, they also take up valuable shelf space. Dissertations may be held in several different media within an institution.

ProQuest Information and Learning, publishing under the UMI imprint, is the designated “national repository” by the Library of Congress, who deems the ProQuest dissertations as a remotely held collection.

Many libraries have been interested in having their retrospective titles placed in this collection since microfilming and digitizing an institution’s dissertations and master’s theses are important ways to showcase its research and academic history, provide access for students and researchers from a single entry point and enhance the institution’s standing in the international academic arena.

In this paper Texas A&M will be used as an example with publisher Austin McLean of ProQuest Information and Learning explaining a case study for a comprehensive dissertation publishing program that focuses on both keeping the data archivally secure as well as increasing access via online distribution.

The paper will detail the library/institution partnership that involved a combination of new and existing services to provide a complete access and archive solution which involved Microfilming. Material Preparation, Target Preparation, Microfilming and Inspection, Negative Storage, Publishing, Scanning (detailing the process by which microfilmed dissertations are loaded onto a SunRise Imaging Proscan 3 models, which are powered by a 300 MHz Pentium P3 processor, 128 MB of RAM, running from a Windows NT 4.0 operating system), Image Storage and Data Sampling / Migration and MARC Records creation.

Background

ProQuest Information and Learning has been publishing PhD and masters theses since 1938 via ProQuest Digital Dissertations/Dissertation Abstracts. Currently the database houses over 2 million citations, with 1.7 million titles available in full text. ProQuest provides theses and dissertations from 1861 to the present in print, electronic and microfilm formats. ProQuest publishes over 55,000 new theses/dissertations each year, with nearly 700 institutions, including every accredited doctoral granting institution in North America, submitting titles to the program.

ProQuest is the designated digital archive for the United States Library of Congress for theses/dissertations. This designation was the first time the Library of Congress has recognized an external site as a repository for a key collection.

Preservation and Access issues with retrospective titles

Since 2000, libraries have been looking for a solution for theses/dissertations that are available in the library on paper without a preservation or access component. These titles predate the university’s submission of titles with ProQuest, having been produced prior to the date the university began submitting titles to ProQuest. In addition, for titles where a microfilm preservation copy has been created, the library desires to have a digital copy made available. Demand for digital access has increased due to the need to serve remote and after hours users.

Libraries have seen the potential for decay, theft, etc, for theses/dissertations, if backup copies are not available. The potential loss would be not only one of scholarship but also part of the institution’s contribution to graduate-level research.

Digital Archiving and Access Program

To meet the needs of the market, ProQuest has created the Digital Archiving and Access Program, which provides a permanent, accessible record of an institution's contribution to higher education. The program consists of the following components:

- Publishing & Preservation
- Enhanced Access

Publishing & Preservation

For a retrospective thesis or dissertation that has not been part of program to date, the title is treated as brand new and is put through the complete publishing process at our headquarters in Ann Arbor, Michigan, USA, including:

- Microfilming Activities
- Material Preparation
- Target Preparation
- Microfilming and Inspection
- Negative Storage
- Publishing in Digital Dissertations Database
- Scanning
- Image Storage and Data Sampling / Migration
- MARC Records
- Perpetual Online hosting

Microfilming Activities

Material Preparation

Material preparation is performed by the Library including complete and accurate collation and securing of loose items and torn pages in the theses prior to sending to ProQuest.

All paper and other materials including photographs, foldout charts, graphs, etc., are microfilmed by ProQuest.

Upon arrival at ProQuest, materials are carefully unpacked by our thesis specialists.

Each title is immediately keyed in to our tracking system and assigned a control number. This number is used to identify and track the status of that title throughout its stay at ProQuest.

All materials are kept on location. The only time they leave the security of our preservation prep room is when they make a short trip to the microfilming and inspection rooms.

When filming is complete, materials are returned to the prep room and shelved according to control number until the film is inspected.

Target Preparation

All target preparation is performed by UMI. There is no charge for standard targets.

Microfilming and Inspection

Up to three generations of silver halide microfilm are produced consisting of a 35mm roll film preservation master negative (first generation), a negative printing master (second generation), and one or more positive use copies (third generation) of each reel (if ordered by the Library).

All filming methods and materials conform to the specific appropriate requirements set forth in the RLG, ANSI and AIIM guidelines for producing preservation quality microfilm.

ProQuest provides 100% technical and bibliographic inspection of the master negative film. Each reel of master negative film is inspected frame by frame for visible defects and missing pages. Second and third generation film is inspected over a light box and on film readers with non-damaging glass plates.

The following specifications are followed when filming a Library's materials:

- All film will be 35mm, nonperforated, silver-gelatin type, on polyester base, as described in ANSI/NAPM IT9.1-1996. Film will be at least 0.13mm (4 mil) thick.
- First-generation film will be Kodak Imagelink microfilm. Second-generation direct duplicating film will be Kodak 2468 or equivalent.
- Processed film will be delivered wound with START target at the outer end, in accordance with ANSI/AIIM MS23-1998, on storage reels that are chemically inert, sturdy, and of dimensions conforming to ANSI IT9.2-1998 and meet the Photographic Activities test for archival permanence.
- Processed film will be stored on reels in boxes made of acid- and lignin-free paperboard that meets the material requirements of ANSI IT9.2-1998. They will be no larger than 4" x 1-5/8" x 3-15/16".
- Once per day a sample of film will be tested for residual thiosulfate using the methylene blue test as described in ANSI/NAPM IT9.1-1996.
- Each roll of first-generation film will be inspected frame by frame by the Library for visible defects and missing pages as described in ANSI/AIIM MS23-1998. Film will be inspected on a microfilm reader.
- Second-generation film will be inspected on a light box to ensure legibility and freedom from defects.
- Every roll of first-generation film will have density readings taken. Average density readings will be established for each reel. It is understood that individual adjustments may be necessary for documents that display variations in paper color and/or ink intensity. Density readings will generally fall between 0.9 and 1.30; however, optimal density will be determined according to ANSI/AIIM MS23-1998, as "that which will make it most legible for reading, scanning, duplicating, or printing to paper."
- All edges of the document will be visible in the image. Reduction ratio changes within the same title will be avoided if possible, but when they must be made, they will be identified by a target in consecutive pagination, whether actually numbered or not.
- Splicing will be kept to an absolute minimum. All retakes will be spliced in proper sequence. All splices will be made with an ultrasonic splicer. There will be no splices in second-generation film. Retakes will include at least the two pages preceding the succeeding pages being refilmed. There will be no splices between the technical target and the text. If the technical target must be refilmed, a minimum of the following ten (10) frames of the text will also be refilmed. Framing will be consistent and regular.

Negative Storage

Storage of master negatives in climate controlled vault located remotely from all other vaults is also part of this turnkey solution.

The master negative vault is 8,300 square feet and contains the original negatives for over 25,000 different periodical titles and in most cases is the complete collection of each title. The archival vault also includes over 1,500,000 these negatives. Over 5.5 billion pages of information are stored within this vault making it the largest commercial library in the world. The film in this vault is for backup and archival purposes only.

Production (duplication) is done from the print masters produced from the originals (see below).

This vault has a separate heating and cooling systems to guarantee constant temperatures of 70 degrees. The vault also includes dehumidifiers that keep humidity between 20% - 30%, adhering to the ANSI standard.

Storage of print master negative (i.e. second generation) in climate controlled vault (separate vault from original camera negative)

The roll film print master vault is 6,800 square feet; the 105mm print master vault is 4,560 square feet. This vault has a separate heating and cooling systems to guarantee constant temperatures of 70 degrees. The archival vault also includes dehumidifiers that keep humidity between 20% - 40%, adhering to the ANSI/PIMA standard.

Publishing in Digital Dissertation Databases

Publishing into the print, CD and online version of Dissertation Abstracts/Digital Dissertations, the world's most used dissertations resource, will entail:

For Bibliographic Data

Bibliographic Control (includes entering title, author, subject, etc. into ProQuest databases)
Publication of bibliographic data in Dissertation Abstracts Database (Electronic, CD, Print versions including DAI (Dissertation Abstracts International), CDI (Comprehensive Dissertations Index), DAO (Dissertations Abstracts On disc), PQDD (ProQuest Digital Dissertations), etc.

For Full Text

Publication in PQDD and Current Research @ (electronic access)
Electronic, Softbound/Hardbound and Microfilm/Microfiche edition available to researchers worldwide.

Scanning

Microfilmed theses are loaded onto a SunRise Imaging Proscan 3 models, which are powered by a 300 MHz Pentium P3 processor, 128 MB of RAM, running from a Windows NT 4.0 operating system.

Scanner operators place the film on the scanners and adjust the image sizing parameters determined by the operators' measurements.

Operators utilize the Proscan 3 image enhancement software to optimize image quality for each title.

Operators key in the reel and catalog number of each work into the proprietary database application that runs on the operator's desktop. The database program accepts only keyed numbers that match an existing

record in the database. This scheme provides a method of fault tolerant redundancy to prevent data entry of incorrect information.

During the scanning stage, the operator's primary objective is to monitor the scan and check for any poor quality images or other inconsistencies such as splices or missing pages.

The primary tool the operator utilizes during the scan is the "image detect" window. Data is output on a graph representation of the microfilm image densities. (The term density is used here in reference to the amount of pixels the scanner senses.) When scanning microfilm with a positive format, the peaks of the graph represent low-density values, and the valleys represent high-density areas. There are spikes in the graph when the scan field crosses a page with an illustration or other dense image. The scanner operator must manually set an average line between the peaks and valleys of the graph while avoiding the data spikes to detect an image.

Software used during the scanning and editing phases runs from Microsoft SQL databases and Visual Basic 5.0 software interfaces developed by UMI's Engineering department.

Operators must visually inspect the microfilm to determine the optimum placement for the scan field to allow for accurate image detection.

Quality control/content development group inspects all scanned images.

Images are viewed using a custom program developed by UMI. The operators approve or reject images based on quality control specifications developed by UMI, with input from focus group research findings.

If necessary, image quality concerns are noted using the custom software. A message is sent to the scanner operator describing any concerns that arises with the images. The scanner operator resolves the concerns and sends the work to be inspected again.

Image Storage and Data Sampling/Migration

Images are stored on a Clarion FC4700 (EMC) which stores 3.5 TB of data. Redundancy is provided through back up magnetic tape.

Adobe PDF format will be the primary archival storage format. All documents will identify file name, file size and creation date. These data will be used for accession to PQDD archival storage.

Sampling Procedures

An automated routine will perform a regularly scheduled sampling activity. The routine will use the creation date or archiving date of active files as a key for sampling. Depending on the actual number of files with the same key date, the routine will sample every 2nd, 3rd, 4th or 5th file. Initially, the to sample will be 20% of the files with the same key date. As digital submissions increase, that percentage may be reduced to maintain a manageable sample population.

Sampling activity will entail scanning files created a year prior to the key date. The same process then takes place for files with a key date two years prior to the start date, etc.,

When discrepancies are discovered, a survey of all files with that key date is triggered. A survey of files in close physical proximity to corrupted files also will take place. If additional discrepancies are uncovered, all files with that key date will be recovered from the reserve storage media and used to replace current files.

Migration and Data Refreshing Policy

All archival equipment and software will have a life of 3-5 years, after which it will be assumed that both software and equipment are obsolete. The media will be refreshed and data migrated as the technology changes on the same 3-5 year cycle.

MARC Records

A US MARC-type record is created for each of the titles scanned or digitized. Fields included in the MARC record include: Author, Title, School, Degree Date, etc. The MARC record can be sent via FTP to the institution for loading in an OPAC. An 856 field can be retroactively added to this record when UMI adds that functionality (target date: 06/04).

Perpetual Online Hosting

All titles are hosted online and made available to the Library through ProQuest Digital Dissertations and Current Research @. No additional fee (annual or otherwise) will be required for online access to these titles (further details are contained in the license agreement).

Enhanced Access

Through the creation of a digital copy, dissertations and master's theses can be made accessible 24 hours a day to users on campus and around the world. The existing microfilm is scanned using the procedure described above.

This program is scalable based on the library's budget and needs. The program can take place by decade, by subject area, by department, by degree type. It can take place over a number of years, or in a single year to maximize one-time funds.

Texas A&M Implementation

Texas A&M chose to digitize the 10,138 titles that were available in the ProQuest microfilm vault for digitization. Reasons cited by the Library for undertaking the project included:

- System-wide coverage for material availability
- One-time cost for project
- Perpetual access to full run via Digital Dissertations
- Reduces time and distance as barriers to dissertation access.
- Increases undergraduate exposure to dissertations.
- Improved coverage for topics that are not indexed in reference materials.

The contract for digitization was signed in November, 2002. Digitization of the 10,138 titles was completed by the Spring '03 semester.

After digitizing the material, the improved access for dissertations was summed up by one Texas A&M reference librarian who said: "(The project) makes dissertations a real information resource." Librarians were also pleased that the project revitalized research by making older dissertations and master's theses easily accessible to a new audience.

Texas A&M is currently considering the second phase of the project, which is to digitize and create microfilm for the approximately 400 titles that are in paper format in the library.

About the Author

Austin McLean is the Director of Dissertation Publishing for ProQuest Information and Learning, Ann Arbor, Michigan, and is responsible for the Dissertations and Master's Theses product lines. Austin oversees a staff who develops and manages Dissertations products in all formats (print, microfilm, electronic). In addition, Austin is working to expand the Publishing program beyond its North American base. Austin received his MA from Northwestern University in Evanston, Illinois, and his BA from DePauw University, Greencastle, Indiana.

U2590