TEXAS TECH UNIVERSITY
Libraries™

DIGITAL PRESERVATION CHALLENGES WITH AN ETD COLLECTION — A CASE STUDY AT TEXAS TECH UNIVERSITY

The Texas Tech community has made this publication openly available. [Please share](#) how this access benefits you. Your story matters to us.

Citation	Perrin, J., Winkler, H., Yang, L. (2014). Digital Preservation Challenges with an ETD Collection: A Case Study at Texas Tech University. Journal of Academic Librarianship, 41;1 Pages 98-104. http://dx.doi.org/10.1016/j.acalib.2014.11.002
Citable Link	http://hdl.handle.net/2346/67413
Terms of Use	CC BY-NC-ND

Digital Preservation Challenges with an ETD Collection – A Case Study at Texas Tech University

Joy M. Perrin

Digital Resources Librarian,
Texas Tech University

Heidi M. Winkler

Digital Curation Librarian Texas
Tech University

Le Yang

Metadata Librarian
Texas Tech University

ABSTRACT

The dangers that digital files face can seem merely the stuff of theory and risk assessment matrices until an institution experiences its first data loss; especially when those digital files represent the graduate research output of a university, the potential impact of that loss increases exponentially. The authors present a case study of the challenges one academic library has encountered in the stewardship of its electronic theses and dissertations (ETDs) over the course of a decade. This article examines the problems that can arise years after the transition from a physical to electronic collection and presents documentation solutions that can make ETD preservation and curation more effective.

KEYWORDS: *Electronic Theses and Dissertations, Digital Stewardship, Digital Curation, Digital Preservation, Dark Archives, Digital Libraries*

INTRODUCTION

In the last twenty years academic institutions have been slowly making the transition from publishing paper theses and dissertations to accepting electronic copies, also known as ETDs. There are many challenges in gathering, accepting, and either publishing or embargoing these electronic copies that have been well established in the literature. One of the challenges that has not received as much attention is that of preserving these electronic documents for the long term. When theses and dissertations were published in paper form, long-term preservation was a simple matter of putting them on a shelf or in a climate-controlled room and recording them in the catalog. Electronic documents require more hands-on curation. This is especially true of ETDs because they come in batches at the ends of school terms, have complicated rights management issues (such as embargoes), and because sometimes there are multiple versions of the documents in multiple management systems. The dangers that digital files face can seem merely the stuff of theory and risk assessment matrices until an organization experiences its first data loss.

The gap in the literature concerning the challenges associated with long-term ETD lifecycle management would indicate that institutions have never experienced data loss, that they have but have not written about it publicly, or that they do not consider the data loss to be substantial. In this article, the case study of one institution's various data losses is presented so that other organizations can learn from these mistakes and secure their data without having to experience the same crises. The lessons learned from these data losses will help any organization looking to preserve not only ETDs but also a wide array of digital files and collections.

LITERATURE REVIEW

Issues concerning electronic theses and dissertations have grown more common in higher education and form important components of today's academic scholarship. Many academic libraries have integrated ETD programs into digital repository and digital scholarship systems (Fox, 1996; Fineman, 2003; Yiotis, 2008; Swain, 2010).

McCutcheon (2011) noted that "the electronic permutations of theses and dissertations had greatly enhanced discovery and dissemination," (p.64) and become the most effective way to distribute graduate-level academic research. In 2004, Eaton indicated that since Virginia Tech implemented the ETD program in 1996, the average print thesis or dissertation had circulated two to three times per year, while in the year 2000 alone, there were an average of over 650 downloads per ETD (in a collection of 5000), coming from the United States and several other countries. Eaton emphasized that "one of the most accessed ETDs had been downloaded tens of thousands of times" (p.6). Lippincott and Lynch (2010) echoed these reports and added that because ETDs have been indexed by major search engines, institutions implementing ETD programs have reported thousands or even tens of thousands of downloads from around the

world. ETDs have become popular among online users because of their easy discoverability, convenient access, and free availability; they thus have become one of the most effective ways for institutions to increase use of theses and dissertations (Fox, McMillan, & Srinivasan, 2012).

Many large research universities have implemented electronic submissions of theses and dissertations through their university libraries. When the submission of ETDs becomes an integral part of an academic system, either mandatory or optional, it represents a complete cultural change in the practice of publishing theses and dissertations extending beyond just formatting issues (Yiotis, 2008). Articles discussing the complexity of ETD programs indicate that successful ETD programs require a variety of collaborative, managerial, and preservation skillsets (Hall, Hoover, & Wolverton, 2003; Jewell, Oldfield, & Reeves, 2006; Fyffe & Welburn, 2008; Yiotis, 2008; Boock & Kunda, 2009; Early & Taber, 2010).

In the survey carried out by the Coalition for Networked Information (CNI) (Lippincott & Lynch, p.9), most respondents reported that libraries, not university faculty or academic administration departments, took the lead in developing and promoting ETD programs and involved some collaboration with the graduate school administration. From the perspective of the graduate school, ETDs raise issues of student training and selecting and maintaining submission systems, while from the perspective of the library, the issues entail publishing and online access, description, and preservation (Yiotis, 2008; Early & Taber, 2010). The roles academic libraries play in the ETD program signify that these institutions should be the entities investing in the long-term curation of these data.

The current literature has frequently discussed issues regarding ETD collaborations, online access software, and metadata (Yiotis, 2008; Lippincott & Lynch, 2010; Early & Taber, 2010; MuCutcheon, 2011). Rare is the literature, however, concerning the long-term digital preservation of ETD programs (McMillan, 2004). Early and Taber noted that as ETD submission has grown in popularity, the focus has shifted from choice of platform to discussions on improving systems. Long-term preservation has also become one of the most important issues that address information storage in the future (Early & Taber, 2010, p.5).

Stewart (2012) observed that many academic libraries had incorporated electronic record systems, like institutional repositories, to handle their born-digital or digitized records, revealing there has been a growing recognition of the need to maintain access to records in electronic format over time. He also discerned, however, that most institutions might have done little to prepare for the long-term preservation of these digital assets outside of these content management systems that were designed primarily for access and easy transactions.

In 2011, Li and Banach conducted a national survey on the digital preservation of institutional repository (IR) materials through ARL member institutions, finding that although 97.4% of surveyed institutions had preservation missions then, only 51.5% had developed policies to support them. The survey also found that many institutions were using IR system features and supplemental limited preservation features, which are not sufficient for long-term preservation purposes. According to Li and Banach, although IR providers had been gradually integrating features for preservation purposes, they could not guarantee the integrity of file

formats for future use. The following case study supports these findings by Li and Banach in tangible real world examples.

CASE STUDY

THE PHYSICAL YEARS (1926-2005)

Texas Tech University (then called Texas Technical College) began producing theses and dissertations in 1929. The physical process of publication changed very little over the next 76 years. At the time, the library handled the binding of these theses and dissertations. The students were required to print out a copy for the library, as well as a copy for the Special Collections Library for archival purposes. Students could also pay to have any number of personal copies bound. Departments, too, would often request a copy. As a result, for each graduating student, there could be anywhere from three to ten copies to bind. The library had staff dedicated to the thesis and dissertation binding and cataloging processes during the months before and after graduation. In the rare case of an embargo, in which the thesis or dissertation would be held back from publication for a set time period, the embargoed thesis or dissertation was simply dropped in the bottom drawer of a filing cabinet in the library until the embargo status expired. A staff person went through the drawer every year and processed any theses or dissertations that had passed their embargo period.

In the early 2000s, the library's administration saw a potential problem: the number of theses and dissertations was increasing from semester to semester, and there was increasingly limited storage for them in the stacks. The library realized that physical theses and dissertations would eventually take over the whole building if a change was not made.

Electronic theses and dissertations were seen as a way to get out of the book binding business and save not only physical space but human resources. Electronic publishing would also be cheaper for the students now that they did not need to have separate copies bound for personal distribution, so it was seen as a move that benefited all involved. Approaching the ETD as a time-saving measure for the library, however, colored how the collection was handled for the next several years. The benefits of open access and the opportunity to engage alumni in different ways were not central reasons for the move but byproducts. The institution would in time learn the consequences of trying to treat ETD curation as a simple, rather than complex, task.

THE FIRST ETD MANAGEMENT SYSTEM: ETD-DB (2005-2009)

Texas Tech University switched to accepting ETDs in 2005 and had success using a system created at Virginia Tech called ETD-db. The ETD-db system gave the university a quick and easy way for students to submit versions of their theses and dissertations electronically, get revisions from the graduate school, and then either embargo their item for a variety of different time periods or publish the documents afterward. The system not only managed ETDs but also

served as the display system for published ETDs. The system allowed students to enter their own metadata, which would then require cleanup by library staff to conform to metadata standards.

Around that time, the library also started a thesis and dissertation scanning effort. The library intended to digitize the 76-year catalog of physical theses and dissertations in order to make them available in the same online location as the new ETDs. The digital lab completed its digitization of the print theses and dissertations in 2009, scanning 12,342 items in total. The scanned items were published in the ETD-db system alongside submitted ETDs.

DATA LOSS #1: LOSS OF METADATA EDITS

In 2006, the library experienced its first major data loss. The database that ran the ETD-db software was corrupted. Library IT staff tried to restore the database through a backup, but they discovered that the backup was corrupted as well. With a 30-day backup, if a problem is not identified in the 30-day window, then the backup is overwritten and the information is lost. Library IT staff then created scripts to re-build the database. All the metadata edits saved in the ETD-db database in that first year were lost in the database reconstruction. The original student-entered metadata was retained, however, meaning that a year's worth of editing work would have to be re-done while new ETDs were still flooding in. This data loss taught the organization three significant lessons.

The first lesson the library learned from this data loss incident was that the systems that are designed for managing and publishing documents are often not well-designed to preserve them. The systems themselves only keep one copy of the documents, and that copy is vulnerable to human administrative error and technical error. If the server had malfunctioned or the database corrupted in a more significant way, the whole system could have been lost. Because the ETD-db system was an all-in-one ETD management and display software, it meant that there was only one copy of items that were put into it. Losing the database or experiencing a problem with the server would have resulted in losing a year's worth of student work and not just the metadata edits but the PDF files as well. The worst case scenario did not happen, but it could have easily happened.

Second, the system backups were not reliable enough to act as a preservation copy. Up until this point, the library had assumed that any data loss would be recoverable by going back to one of the 30-day backups at the central campus data center. IT backup copies are not designed to be permanent. They change periodically and are vulnerable to corruption or other errors like this backup was.

Finally, those involved realized that digital items are collections of files. The library had been focused on preserving the ETD PDF files themselves by creating checksums (numbers created by an algorithm to detect file error) for them but had neglected to take steps to preserve their metadata.

DATA LOSS #2: LOSS OF ADMINISTRATIVE METADATA-EMBARGO PERIODS

In 2009, the library transferred the ETD collection from the ETD-db system to DSpace to accommodate the use of new, separate grant-funded ETD management software that was being adopted by most of the academic institutions in the state of Texas. This move was coordinated between a metadata librarian and a programmer from the library's IT department. All the files from the ETD-db system were extracted. These files included all the documents that had been scanned or submitted and all the metadata that had been created. The files that were openly available in ETD-db were imported into DSpace. The embargoed files were not moved because of questions of how to display them.

This move nearly caused a second preservation disaster. Because the embargoed items did not get published into the new system, there was no official location for those items to be held. If there were plans for where to place these items other than DSpace and how to handle the expiring embargo periods, those plans were not documented and the people involved left the institution before the full extent of the preservation crisis was realized. Between 2009 and 2011, the embargoed items from 2005 to 2009 were effectively lost.

Years later, a zipped file was found on a server that had been set aside for temporary data dumps. After extracting the zipped files, it was discovered that this file contained the original exported files from ETD-db before it was processed to be ingested into DSpace. This meant that it contained not just the published items but also the embargoed items that had not been published. This zipped file became an accidental archive of all of the digital items embargoed in the ETD-db System.

This accidental archive contained 17,000 folders of PDFs and metadata files of both the scanned and born-digital ETDs. It was moved to a working server with the other scanned files, meaning that for those scanned items that had been published in ETD-db, there were now duplicates of those files on the server. The library did not have the time or resources to consolidate and organize the two groups of files, so they decided to just hold both copies since server storage space was plentiful and cheap.

The library was fortunate to have found the items before they were lost, but the way the items were saved did not preserve the information on embargo periods for the theses and dissertations that were not published (e.g., one year, two years, or permanently) because this data had been held in the ETD-db system's database separate from the metadata or the PDFs. Because some ETDs were embargoed due to nondisclosure agreements or defense contracts, and the time period on those embargoes was vital for the student graduating to maintain their agreements with other entities, the library staff involved decided to assume that if it was not available publically on DSpace, then the item was permanently embargoed.

Not unlike the accidental archive of the files, a chance conversation while the authors were gathering information for this article led to the discovery of an informal copy of the ETD-db database that had been made by one of the library IT staff who had worked on the project. He said it was his practice to make a copy of every database he worked on before he worked on it, and the database files were sitting in a personal drive. The ETD embargo statuses that had been

missing for five years were recovered; however, they could have easily been lost forever. Something could have happened to the drive they were on, the database files could have been corrupted, or the knowledge that the data were there could have simply never come to light. It would be better and more secure if the data were not preserved by chance but by careful design and policies.

Losing the embargo periods was a disservice to the graduate students who had expected their work to be published after a set period of time. The library learned that this crisis could have been avoided if the organization had kept a separate record of the embargo status of the items outside of the management system.

DATA LOSS #3: LOSS OF SCANNED FILES

In 2009, the library had one server dedicated to digital files. This server was both the working server and the location to store files after scanning. This working server had a 30-day backup. At the time, the library felt confident that 30 days were more than adequate to ensure items were safe from corruption or accidental deletion before being cleared from the server. This assumption was proven wrong quickly. The staff believed that the 30-day backup was sufficient partly because the incident with the first data loss was not recorded and shared. The lessons learned from Data Loss #1 that did not get passed down or recorded.

Human error resulted in the deletion of over a thousand scanned ETDs. Once the error was discovered, the staff attempted to go to the 30-day back up and retrieve the deleted files only to discover that the accidental deletion had occurred more than thirty days before the backup. Much like the database corruption in the ETD-db system, the mistake had happened beyond the 30-day window that the backups were good for. The library had not learned from Data Loss #1 and planned for an event where a data loss would not be found within the 30-day backup cycle, and it had also not put into place policies that would have helped prevent a human administrative error.

This data loss was recoverable. The files that had been deleted were all published, so the published copies were downloaded and put back onto the server. The data loss did, however, highlight the need to change policies and create a better framework for preservation. A separate server space was set up for archive items so that they could be kept separate from working data. This is how the libraries' dark archive started.

DATA LOSS #4: FAILURE OF POLICIES – LOSS OF EMBARGO STATUSES CHANGES

Between 2009 and 2012, the new ETD management system was working well. It pushed published items to DSpace and held back embargoed items. Something unexpected happened along the way, though. Library staff assumed at first that graduating students would understand the consequences of both embargoes and publishing and would know enough to decide at the point of submission what they wanted for their thesis or dissertation. That assumption was

proved wrong as students started calling the library to have their status changed soon after graduation. Students who had embargoed their work wanted to publish the work so that employers could read it. Students who had published their work wanted the items taken down so they could get the work published through more traditional means. While the library was able to accommodate the embargo status changes, the individuals involved failed to record the changes in a cohesive and organized way. The end result was that the embargo statement that was recorded in the ETD management system did not match what was published in DSpace. For items that had been embargoed after being published, the records were simply deleted out of DSpace, and a new embargo period was not identified.

As before with the ETD-db system items, library staff must now assume that all items that were not published in DSpace are permanently embargoed. This could have been avoided by learning the lesson from Data Loss #2 and implementing a separate location to preserve embargo statuses.

THE START OF DIGITAL PRESERVATION (2011-2012)

In 2011, the library started on a path toward real digital preservation for the ETD collection. Funds were set aside to increase the storage of the archive space to prepare for future storage needs as well as to purchase another server to act as a second copy of the archived files. The second archive server was installed in a building twenty miles outside of town, giving the library two copies of the archived files in geographically distant locations. Not only was there now a second copy of everything in case of human or technical error, but in the event that a disaster took out the on-campus servers (such as a tornado or flood to the central campus IT building), there would still be a copy of the files from which to recover. The library also implemented a daily checksum check on the two different copies. If one copy of the archived files changed or got corrupted, there would be a notification the day after it happened and the other copy could be used to recover the file.

At this point, the ETDs could be considered to be preserved at a bit level. This meant that the data itself had a very low chance of being lost even with human error. In addition to preserving the bits and bytes of the files, a central document was created to identify each item and its embargo status outside of the ETD management system. This file was called the manifest and was put on the archive servers to be preserved in two copies in two locations along with the files.

The library took advantage of this bit-level preservation by quickly moving all digital files to the servers, even if the files were not organized. The idea was to preserve them so that the files would not be lost, even if there was overlap between the groups of files. Not only was there overlap, but each group of files had its own organizational system. Whereas one group was organized under name, another group was organized under system numbers, while another group was organized under date submitted. As a result finding any ETD in the archive was difficult. Someone would have to know that there were many different places to look and to understand each organizational system in order to see if a particular thesis or dissertation was in the archive

at all. The only way to organize these documents would be to meticulously identify them one-by-one.

THE PATH TO RECOVERY (2012)

In 2012, the library formally created a digital resources unit to handle the ETDs and other digital initiatives. The creation of this unit finally brought the entire digital workflow, from scanning to preservation, under one supervisor. They soon started getting calls from students who could not find their theses or dissertations online as they had expected. Each request involved sorting through the three or four different groups of files with overlap containing three to four different organization systems. Staff realized that these searches were taking up a lot of time and that it would be easier if there was one place to archive ETDs under one organizational system.

A universal naming system was created for this future single ETD archive. The ETD collection was given an overall name of “etd001”, which identified it as the first ETD archive with only six characters. Each item was given a six-character identification number and a folder for all the files related to that item. The first group of files that were organized into this system were the files that were published in DSpace. The digital items containing both the PDFs and the metadata were preserved under this naming system. The manifest was updated with identification numbers, author names, titles, the year the student graduated, embargo status if it was known, and a Uniform Resource Identifier (URI) or the item if the item was published. Even though the ETD archive was not complete, the organizational system and manifest made finding ETDs much faster. If the embargo status of an ETD was changed, a note was added to the item on the manifest and a copy of the correspondence was saved in the item’s folder in the archive.

Library staff started going through the other groups of items. The first step to moving all the groups into one archive was redoing the organizational systems so that they would all be organized in a way that was both easy to search and easy to integrate into the main archive. The staff decided to organize all the other ETD groups by author name so that duplicates would be more visible, and the lost ETDs would be easier to find if the author’s name was known. By undertaking this renaming, the unit was able to find all of the previously “lost” theses and dissertations. The unit was on the path to recovery, but there was still a lot of work to be done to clean up the collection. There was still a significant overlap of copied items between the different groups. At the time, this was not a problem since the library already had plenty of purchased storage. This, however, changed when the library decided to work toward long-term digital preservation.

The Digital Preservation Network (DPN) was just starting as an idea in 2012. DPN’s stated goal is to gather academic institutions together to provide a location to preserve academic digital content for long-term preservation. The network was sold as a way for an organization to have multiple copies of their digital files in different locations around the globe and on different technical architectures. The network’s goal would be to preserve the files long after an institution

ceases to exist. The library joined DPN in hopes that it would yield a level of preservation far beyond what the university would be able to accomplish alone.

DIGITAL STEWARDSHIP (2013-PRESENT)

In order to support the transition from solely local digital preservation to a future with complementary long-term preservation through DPN, the library hired a digital stewardship librarian. This librarian was put in charge of the archive and tasked with cleaning up the archived digital collections in preparation for future long-term preservation. With every terabyte of information that might go into DPN incurring a possible yearly fee for preservation, it became important to not have redundant files preserved on the same server. A focused effort was then put toward going through all the copies of the files and moving them all into the official “etd001” archive with no overlaps. To start this process, the digital stewardship librarian would need to identify every ETD that had gone through the library. Everything that was supposed to be in the collections needed to be accounted for; all the duplicates and any other items that should not be in the collections needed to be weeded out, and all of this needed to be properly organized. It is estimated that, as of 2014, the ETDs alone will take two years to audit.

CURATION SITUATION (2014 AND ONWARD)

Figure 1 is a visual representation (not to scale) of the different ETD groups that came to be through the series of system changes and data losses described in detail above. The places where the circles overlap represents files duplicated. For example, all the files in DSpace are in the etd001 archive, but not all of the ETD-db files are in the etd001 archive.

Figure 1: Groups of ETD Copies and How They Overlap (not to scale)

The etd001 archive is the folder in the dark archive where the digital resources unit intends to house all the ETDs both published online and embargoed from public view. The DSpace instance is the library's official published archive, displaying all published ETDs and records for embargoed ETDs.

The process of curating the collection will entail making sure that items are represented in the public DSpace and the dark archive of etd001. This ideal situation is represented by Figure 2.

Figure 2: The Ideal Situation for ETDs (not to scale)

It is estimated that the process of moving the archive from where it currently is to that of the ideal situation represented in Figure 2 will take a total of two years.

A report of all the files in each group was made and then combined into a master list that included the location of the file (which of the different groups the file was in) and the name of the student on the file. This was only possible because of the renaming effort that happened in 2012 where each file was looked at and re-named to represent the first and last names of each student. The list was then sorted by last name. This makes identifying duplicates much easier. The digital stewardship librarian's first task in this audit was identifying which copy to keep and which copies are duplicates. The next step has been ensuring that every ETD represented online or within the original graduation lists is also in the archive. A sample of this report that the audit is being conducted off of is shown in Figure 3.

	TITLE	DEPARTMENT	DEGREE	ARCHIVE LOCATION	EMBARGO	URI
	Exploring Motivational Factors in College Students' Foreign Lan Education	Psychology	Doctoral	TEMPORARY COPY	VT: Withheld	
	Eeg Power and Coherence Features of Alzheimer's Disease and Electrical Engineering		Doctoral	TEMPORARY COPY	VT: Withheld	
	Automatic and Interactive Texture Segmentation Techniques fr Electrical Engineering		Doctoral	TEMPORARY COPY	VT: Withheld	
	Ernest Hemingway and Gabriel Garcia Marquez: Cultural Ascenc	English	Doctoral	???		
	The Gothic Short Stories of Louisa May Alcott in the Context of l English		Doctoral	???		
	Development of Resistance to Agrochemical in Thamnocephali	Environmental Toxicology	Doctoral	TEMPORARY COPY	VT: Restricted (On campus only)	
	Cognitive Impairment in Mdx Mouse	Experimental Psychology	Doctoral	???		
	Judgements of Collision: Theoretical Practical Implications	Experimental Psychology	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Behavioral Complexity and Effectiveness Among Cooperative E Family & Cons Sci Educ	Doctoral	Doctoral	TEMPORARY COPY	unknown	
	Family and Consumer Sciences Teachers' Special Education Nee Family & Cons Sci Educ	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	
	A Primer on the Aesthetics, Fabrication, and Culture of Lowride Fine Arts (Art)	Doctoral	Doctoral	ttu_etd001_012365, TEM	unknown	http://hdl.handle.n
	Fine Arts Faculty Evaluation Process in Higher Education: Its Val Fine Arts (Theatre)	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Effects of Cadmium at Low Concentrations on Anuran Developr Fisheries Science	Doctoral	Doctoral	TEMPORARY COPY	VT: Withheld	
re	Petrology and Emplacement History of the S-Type Intrusion, Si Geoscience	Doctoral	Doctoral	???	unrestricted	http://hdl.handle.n
	Atmospheric Boundary Layer Evening Transitions Over West Te Geoscience	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Stable Isotope Geochemistry of Upper Cretaceous and Paleoccei Geoscience	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Assessment Training for Faculty: Effects on K-12 Teacher Prepai Higher Education	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Faculty Actions and Student Satisfaction in Distance Education Higher Education	Doctoral	Doctoral	TEMPORARY COPY	VT: Withheld	
	Understanding the Generation Y American Food Market: the Ro Hospitality Administration	Doctoral	Doctoral	TEMPORARY COPY	VT: Withheld	
	Negotiating Identity: a Qualitative Study of Second Generation Human Devlmt. & Family	Doctoral	Doctoral	TEMPORARY COPY	VT: Withheld	
	Selection and Socialization of Young Adult Romantic Partners' C Human Devlmt. & Family	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Simulation of Tossing - Biomechanics and Motor Control	Industrial Engineering	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	Integrated Approach to Vehicular Movement Network Modelin Land Use Planning Mgt & D	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
	An Adolescent Behavioral Health Screening, Distinct Primary C Marriage and Family Thera	Doctoral	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n
her	Parental Influence on Hispanic Adolescent Substance Use	Marriage and Family Thera	Doctoral	TEMPORARY COPY	unrestricted	http://hdl.handle.n

Figure 3: Screen Shot of the ETD Audit

As an example of how flawed the organization of the ETD collection had been, just a few weeks into the audit, the digital stewardship librarian found one scanned thesis that had two archive records and four DSpace records. The fourth DSpace record pointed to an additional newly scanned thesis located in the folder where items sit before final archive processing. That meant there were seven copies of one thesis. These kinds of unexpected problems are a testament to how difficult the clean-up for the ETD collection has been. Going through the entire collection and getting rid of these kinds of problems will take years and could have been avoided if the institution had implemented proper records-keeping and kept the files organized in one archive from the very beginning.

CONCLUSION

Many of the problems that this ETD collection encountered could have been avoided by following simple digital preservation and documentation policies. Other institutions seeking to avoid similar crises would do well to pursue the practices that follow.

Institutions must make digital preservation plans beyond backups that are only temporarily accessible. Traditional backups are not enough to protect the digital files because they are designed to be short-term solutions. The library learned that if data loss happens outside the range of the backups, the data are lost forever. In addition, the backups themselves might fail. Either new information is not captured, or the backup is unable to be retrieved. Institutions should develop processes that will allow the institution to keep files for decades rather than keeping working data for only a few weeks. True digital preservation requires active management on systems designed to hold information for much longer periods of time.

Systems primarily designed to publish electronic files are not preservation solutions. These systems often have small features that help with some preservation, but they are not designed to preserve the files for the long term. Organizations with valuable digital assets should make the investment in their items to store them outside of any display system they might have. Institutions may be tempted to view their one system as a total access and preservation solution, but these content management systems made for displaying digital content are not ideal for preserving long term. Once the library accepted that, it was able to move forward with purchasing external storage and making the effort to organize the files that were kept on these servers.

Two to three different copies of digital items must reside on different servers in order to guarantee that files will not be accidentally deleted or lost through technical difficulties. These copies need to be organized and identical. While the library did end up with duplicate copies for some digital items, they were unorganized and all on the same servers. In order to be truly preserved, the items need to be cohesively organized and in two to three different places that are physically separate. This is the first step in any digital preservation process. Institutions should implement this practice as soon as possible after beginning their ETD programs so that they do not have to play catch-up later on down the line, post-digital preservation crisis. This second copy does not have to be across the world. Any amount of space between the copies is better than none at all.

All metadata, to include administrative data such as embargo periods, should be preserved outside of the display system. The metadata is a crucial part of the digital item. Institutions can expect much more work to recover lost metadata than to recover a lost digital item, but in most content management systems the metadata is not as protected as the item. A single list of ETDs and their embargo statuses (and links to the item online) is extremely valuable. Had the library been keeping this list from the very beginning, the digital resources unit would not be spending years creating the list and would instead have more information about items that were embargoed after being published and why.

Maintaining written procedures and documentation for all aspects of digital collections is vital. In many ways, preserving institutional memory and the reasons why decisions are made are as important as preserving digital assets. In 2013, the library implemented an archive log where changes in policies and all additions, deletions, and moves are recorded. This allows the staff to track the collections and understand what was done and why. Having well-organized multiple copies of finished items outside of the system the items were published in would have prevented the initial loss of metadata edits and the later loss of scanned files. Having a record of the administrative metadata outside of the content management systems would have saved the lost embargo information in a purposeful way instead of relying on chance copies.

The success of digital preservation will require collaboration between the curators of these collections and the IT resources who maintain the software and hardware. This collaboration may prove difficult simply because of differing professional cultures. The library world's understanding of digital preservation lasting for decades or more often clashes with the IT world's expectation for files only to survive for a short time. In this case study, the word

“archive” did not have the same meaning for someone with a library science background as it did for someone with an information technology background. If a librarian would ask the IT staff shepherding the ETD collection if the items were archived, the answer was always yes. From an IT perspective, however, archiving the items meant backing them up for 30 days or publishing them to the online archive. Throughout the lifecycle of the ETD collection, these kinds of language differences between the library and IT proved to be a challenge. It took a number of years to communicate the need for extra servers in separate locations and to make a case for the amount of storage needed. To the IT people involved in the beginning, these measures seemed to be costly and pointless. It was hard for them to see why more of the budget needed to be spent on servers that, from their perspective, would be doing nothing. Not until the IT department saw the wider ramifications of losing this data did they truly understand why the library needed to invest in keeping multiple copies in different locations. Librarians need to be adept at communicating with IT by speaking in their terms and explaining the importance of the data being preserved.

A final, major factor in the library’s recovery and creation of preservation and curation strategies was having an internal digital resources unit dedicated to ETD management. Before, the staff involved worked on ETDs in addition to their other responsibilities, which caused a lack of comprehensive focus and meant there was not enough time to spend on cleaning up problems and organizing items. Communication across departments was difficult, and any staff turnover meant a loss of institutional memory. This new unit had the agility to devote more human resources to ETDs when necessary, and the unit could argue effectively for money for servers and greater digital preservation solutions such as DPN. The unit could also keep track of known weaknesses with the data, plan for those weaknesses, document the whole process, and ensure that the documentation was available even with staff turnover. The library learned that digital collections deserve formal curators. If the collection is written into someone’s job description, then that collection has a better chance at survival and digital preservation disasters or data loss will be noticed more quickly.

Ultimately, every ETD program is different for every institution. Some programs accept a wider variety of file formats while others have stricter embargo policies. At some institutions, the library does not play the same major role in ETD publishing as it does at Texas Tech University. Many schools, though, will admit to having implemented ETD programs before they understood the basics of digital preservation. Everyone associated with an ETD program is learning as they go, and what is equally important to having policies that anticipate digital preservation disasters is learning how to react to the disasters when they do happen. The first step in recovering from a digital preservation failure is not to panic. Instead, talk to everyone involved, including IT for any possible chance that an otherwise unknown copy of the data exists. Have a backup plan for data recovery. For Texas Tech University, the Graduate School had a comprehensive list of graduates that allowed those involved with the loss to decide what data was missing. Any ETD program can benefit from brainstorming about various data loss scenarios and how the data could be recreated or reconstructed.

Even though this case study has primarily been a description of local issues, the grander lessons gleaned from these crises are not specific to this institution. Librarians are learning and

re-learning every day that digital collections cannot be managed in the same fashion as their physical counterparts. These digital collections require more active care over the course of their lifecycles and may require assistance from those outside the traditional library sphere, such as the graduate school or an IT department. As shown in the case study, preservation efforts will be needed at different times in the lifecycle of a collection. These moments demand active curation from someone looking at the whole collection with an eye toward its long-term survival. An active, concerned curator may be all a collection needs in order to survive without major data loss.

While these particular problems happened with an ETD collection, they can very easily happen to other kinds of digital content in other situations. With more and more academic research being produced digitally, and with open access mandates and data management becoming more important to researchers, academic libraries need to come up to speed about digital preservation quickly to avoid unintentional data loss of important academic works.

REFERENCES

- Boock, M., & Kunda, S. (2009). Electronic thesis and dissertation metadata workflow at Oregon State University Libraries. *Cataloging & Classification Quarterly*, 47 (3-4), 297-308.
- Early, M. G., & Taber, A. M. (2010). Evolving in collaboration: Electronic thesis and dissertation workflows in North Carolina. *Collaborative Librarianship* 2 (1), 4-18.
- Eaton, J. (2004). Enhancing graduate education through electronic theses and dissertations. In E. A. Fox (Ed.), *Electronic theses and dissertations: a sourcebook for educators, students, and librarians* (pp. 1-7). New York: Dekker.
- Fineman, Y. (2003). Electronic theses and dissertations. *Portal: Libraries and the Academy*, 3 (2), 219-227.
- Fox, E. A., Eaton, J., & McMillan, G. (1996). *Improving graduate education with a national digital library of theses and dissertations: proposal submitted to the United States Department of Education*. Retrieved July 15, 2013, from <http://www.ndltd.org/support/fipse10.pdf>
- Fox, E. A., McMillan, G., & Srinivasan, V. (2012). Electronic theses and dissertations, progress, issues, and prospects. In T. W. Luke and J. Hunsinger (Ed.). *Putting knowledge to work and letting information play*, (pp. 95-110).
- Fyffe, R., & Welburn, W. C. (2008). ETDs, scholarly communication, and campus collaboration. *College & Research Libraries News*, 69 (3), 152-155.
- Hall, S., Hoover, L., & Wolverton, R. E. Jr. (2003). Publishing electronic theses and dissertations: reconfiguring library services. *Technical Services Quarterly*, 21 (2), 63-70.
- Jewell, C., Oldfield, W., & Reeves, S. (2006). University of Waterloo electronic theses: Issues and partnerships. *Library Hi Tech*, 24 (2), 183-196.

- Li, Y., & Banach, M. (2011). Institutional repositories and digital preservation: Assessing current practices at research libraries. Retrieved July 15, 2013, from <http://www.dlib.org/dlib/may11/yuanli/05yuanli.html>
- Lippincott, J. K., & Lynch, C. A. (2010). ETDs and graduate education: programs and prospects. *Research Library Issues*, June 2010 (270), 6-15.
- McCutcheon, S. (2011). Basic, fuller, fullest: Treatment options for electronic theses and dissertations. *Library Collection, Acquisitions, & Technical Services* 35 (2011), 64-68.
- McMillan, G. (2004). Digital preservation of theses and dissertations through collaboration. *Resources Sharing & Information Networks*, 17 (1-2), 159-174.
- Stewart, C. (2012). Preservation and access in an age of e-science and electronic records: Sharing the problem and discovering common solutions. *Journal of Library Administration*, 52 (3-4), 265-278
- Swain, D. K. (2010). Global adoption of electronic theses and dissertations. Retrieved July 15, 2013 from <http://digitalcommons.unl.edu/libphilprac/418>
- Yiotis, K. (2008). Electronic theses and dissertation repositories: What are they? Where do they come from? How do they work? *OCLC Systems & Services*, 24 (2), 101-115.