Stewardship of Electronic Theses and Dissertations: Risks and Costs

Lorraine L. Richards

2017-2018 Drexel University Libraries' Library Fellowship Report

Stewardship of ETDs: Risks and Costs

Summary

In April, 2018 Drexel University Libraries finalized a movement from the storage of paper-based electronic theses and dissertations (ETDs) as copies of record to an electronic option. Now, the ETDs, while still being registered with ProQuest for electronic access to the wider world, are stored onsite, using Drexel's College of Computing and Informatics' Technology Department. Although such activities have taken place by some libraries since at least the mid-1990s (Weisser and Walker 1997), many libraries have found that using commercial ETD services provide the discovery mechanism they desire and retain hard-bound copies of theses and dissertations within their archives. This has been the approach taken by Drexel Libraries until now.

This paper represents the results of a Memorandum of Understanding (MOU) between the author and Drexel Libraries to examine the ongoing electronic storage options being undertaken for the Drexel ETDS as a source of record for Drexel theses and dissertations, examining the risks to their "trustworthiness" as records and the costs of maintaining the records onsite. The primary charge of the MOU involves the following components:

1. Develop understanding of the state of the practice in archiving electronic rather than print theses and dissertations among US academic libraries.
2. Commentary and if needed, proposed clarification in documentation of the Libraries' stated service change and planned procedures and arrangements with ProQuest, and with submission/ingest workflows into iDEA (i.e., Drexel's onsite storage source for ETDs).
3. Review of the Libraries' assessment of IDEA as a level of "trusted archive" and conduct additional assessment, if needed, sufficient to maintain an archive of electronic theses and dissertations to meet records management responsibilities, and the level of risk to rely upon ProQuest as an external vendor to supplement the role at this time. Report of the review and recommendations for priorities to continue to improve the local repository.
4. Provide evidence and recommend arguments to address concerns about reliance on electronic copies for discoverability, access, and long term availability of theses and dissertations.

5. Estimate of costs to maintain the archive of theses and dissertations with a high level assessment of savings [for students and for the University] generated by conversion to an electronic archive

A series of interviews were conducted with Danuta Nitecki, Deb Morley, John McNamara, Steve Aucott, David Cupo, and Matthew Lyons, with data provided by Ann Yurcaba. All are important players in the ETD implementation.

State of the Practice of Maintaining ETDs

Numerous libraries worldwide have taken up the use of ETDs in the past twenty-some years. An early adopter, Virginia Tech, began requiring their students to submit an electronic copy of their dissertation in 1997 (Yiotis 2008). Drexel has likewise required electronic submissions within the past decade, but has maintained the versions of record in hard-copy form, stored in the University Library.

The primary difficulty most libraries face in implementing ETDs is political and cultural. Because Drexel has already an established ETD program which feeds ETDs directly to ProQuest, this risk is less than if they were beginning from scratch. Nonetheless, continued communication with graduate programs and schools is still of ongoing concern, and changes to processes should be communicated in a clear and transparent way, focusing on the reasons for changes and the benefits and risks of those changes. For example, per Deb Morley, former Director of Data & Digital Stewardship at Drexel Libraries, to eliminate the print archival copy of theses and dissertations, the graduate schools will no longer need to have the Library's signature on the completion form, while the approval form must be performed electronically. A request process had to be created to process embargoes. Additional changes to the process, such as moving storage from iDEA to DuraSpace, should be carefully conveyed to all stakeholders, along with the implications of these changes.

Maintaining ETDS, when done electronically, is typically either performed by the library itself, through their computer operations department, or in the case in which such a department doesn't exist, through a computer service resident at the University but separate from the library. Some libraries maintain their ETDs in their digital libraries (e.g., the California Digital Library) or through an institutional repository. A relevant source provided by the Networked Digital Library of Theses and Dissertations (NDTLD) provides an "ETD Guide" that walks one

through the necessary steps of setting up an ETD repository from scratch (http://etdguide.ndltd.org/).

Nonetheless, more recent efforts to set up ETD systems have focused on the ever-important question of trust. A trustworthy repository ensures the long-term safety, access, discoverability, and security of the digital information retained within it. Drexel has taken the stance that trustworthiness of the iDEA repository is of utmost consideration before moving to its planned preservation strategy and operations when greater resources are available.

The attributes of a trustworthy digital repository include:

- Compliance with the Reference Model for an Open Archival Information System (OAIS-RM)
- Administrative Responsibility
- Organizational Viability
- Financial Sustainability
- Technological and Procedural Suitability
- System Security
- Procedural Accountability

Although these requirements are in themselves sometimes onerous, what many do not realize is that simply achieving these goals is not enough to prove trustworthiness. Rather, each of these goals requires extensive documentation as to how they are defined and how they are met, such that an individual examining the preservation practices of the organization can assess the organization's compliance. At this point, some documentation regarding financial sustainability is available, in the form of an MOU with the CCI Information Technology department, but other documentation is either missing or not yet compiled in a readily-discoverable format. A strong recommendation of this report is that such documentation begin being developed while Drexel Libraries prepares its move to DuraSpace. Although this will not grant it a certification of trustworthiness (a long and expensive activity), it will be able to declare that it has gone through its own self-assessment of trustworthiness according to TRAC guidelines, or at least the NDSA Levels of Preservation (Owens, 2012) .[1] These levels can be seen in Figure 1. This would place Drexel Libraries in a superior position with respect to trustworthiness, in comparison with many libraries, both larger and of similar size.

[1] Trustworthy Repositories Audit & Certification (TRAC) is a document describing the metrics of an OAIS-compliant digital repository that developed from work done by the OCLC/RLG Programs and National Archives and Records Administration (NARA) task force initiative.

Table 1 - NDSA Levels of Preservation

	Level One (Protect Your Data)	**Level Two (Know Your data)**	**Level Three (Monitor Your Data)**	**Level Four (Repair Your Data)**
Storage and Geographic Location	Two complete copies that are not collocated For data on heterogeneous media (optical disks, hard drives, etc.) get the content off the medium and into your storage system	At least three complete copies At least one copy in a different geographic location Document your storage system(s) and storage media and what you need to use them	At least one copy in a geographic location with a different disaster threat Obsolescence monitoring process for your storage system(s) and media	At least 3 copies in geographic locations with different disaster threats. Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	Check file fixity on ingest if it has been provided with the content Create fixity info if it wasn't provided with the content	Check fixity on all ingestsUse write-blockers when working with original media Virus-check high risk content	Check fixity of content at fixed intervals Maintain logs of fixity info; supply audit on demand Ability to detect corrupt data Virus-check all content	Check fixity of all content in response to specific events or activities Ability to replace/repair corrupted data Ensure no one person has write access to all copies
Information Security	Identify who has read, write, move, and delete authorization to individual files Restrict who has those authorizations to individual files	Document access restrictions for content	Maintain logs of who performed what actions on files, including deletions and preservation actions	Perform audit of logs
Metadata	Inventory of content and its storage location Ensure backup and non-collocation of inventory	Store administrative metadata Store transformative metadata and log events	Store standard technical and descriptive metadata	Store standard preservation metadata
File Formats	When you can give input into the creation of digital files encourage use of a limited set of known	Inventory of file formats in use	Monitor file format obsolescence issues	Perform format migrations, emulation and similar activities as needed

	open file formats and codecs			

Drexel Libraries' workflows are highly granular compared to many of those currently available to the public. Most published examples simply discuss the process the student follows until they upload their final approved thesis or dissertation to ProQuest or their institution's archive. Others provide high level diagrams, such as University of Oregon's workflow for their born digital and converted-to-digital ETDs. (At this point in time, Drexel Libraries will maintain their legacy, hard-copy theses and dissertations in hard-copy format.

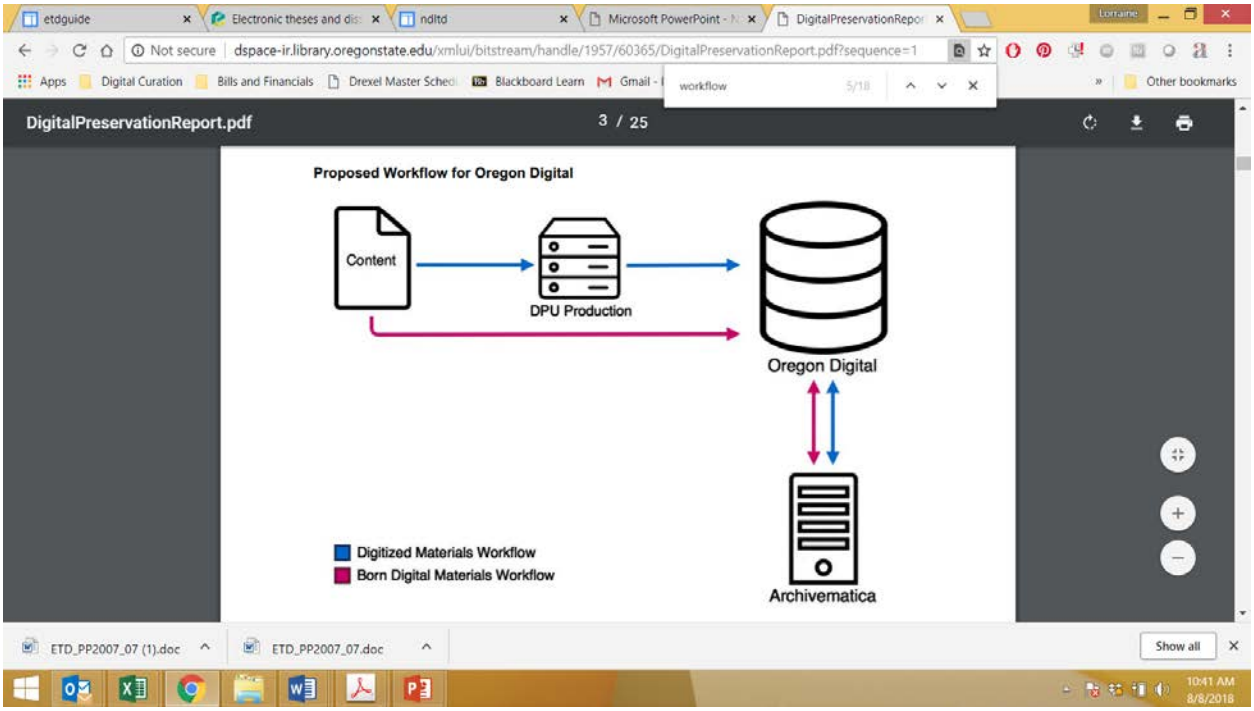

Figure 1 - University of Oregon ETD Workflow

As one can see, this is not a comprehensive picture of activities. Other discovered workflows available publicly are of similar generality. For example, Ohio State University provides the following workflow:

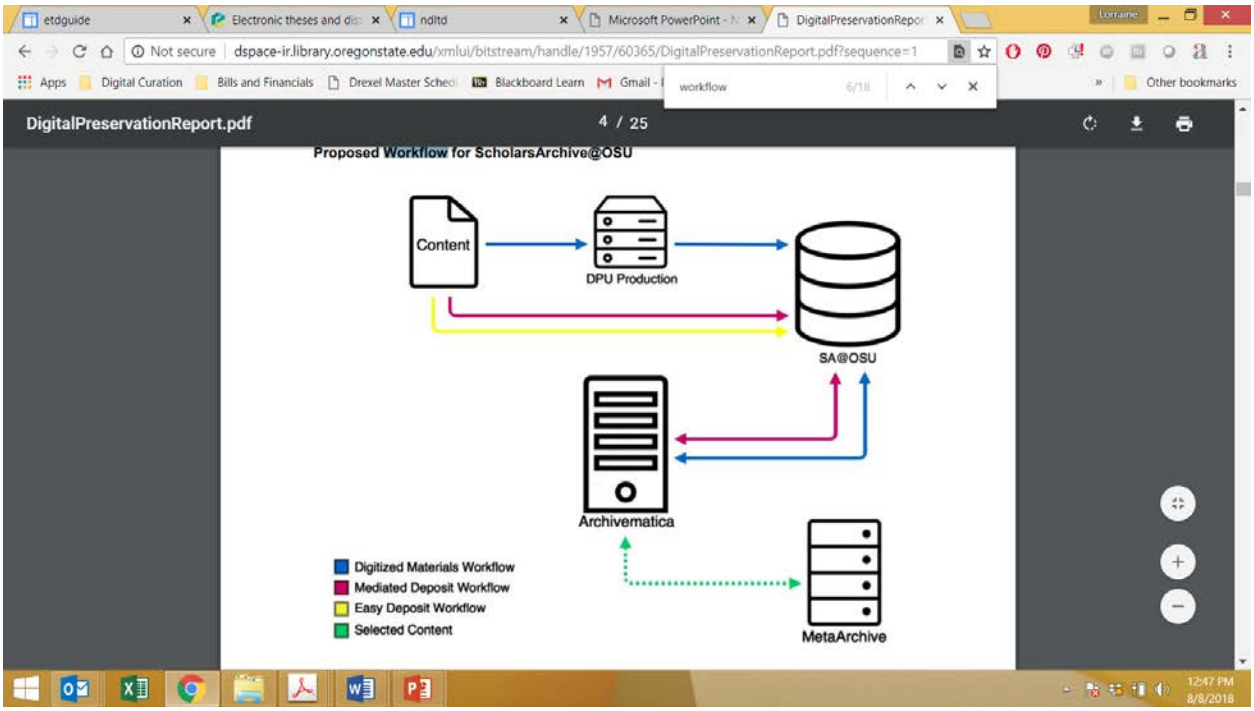

Figure 2 - Ohio State University ETD Workflow

Current available workflows focus upon the flow of data only and do not include internal operations that correspond to the flows. Drexel Libraries, however, has depicted their workflow in much more detail, as seen on the next page in Figure 3. Drexel's detailed workflow presents an improvement over such diagrams, insofar as it links the internal, operational activities associated with its preservation and access actions with the technical events involved in supporting those activities.

Currently, this chart is not entirely accurate, but rather, is a depiction of the future of the ETD workflow. In the lower right hand corner, "DuraCloud" is depicted as a resting place for ETDs. In fact, due to resource shortages, that particular portion of the flow is not yet in place. Rather, ETDs find their final resting place in iDEA, the Drexel-developed system, hosted by the College of Computing and Informatics Information Technology Department.

Description of the Workflow

Internally, the workflow begins with the student who wishes to submit his or her theses or dissertation (T/D) and his or her faculty advisor. The student fills out an online approval form, along with the T/D and submits it. It is placed in a staging area (ETD ADMIN) prior to being uploaded to ProQuest. A verification is created automatically and stored for the availability of the Graduates College and Graduate School of Biomedical Sciences and Professional Studies. If the student desires an embargo, the faculty advisor

must approve the embargo ad and embargo request is submitted electronically to the University Archives. The request is emailed to iDEA, so that when the upload of the T/D occurs, it is flagged as unavailable for access during the embargo period. (The ProQuest submission also includes an embargo request.)

Description of the Workflow

Internally, the workflow begins with the student who wishes to submit his or her theses or dissertation (T/D) and his or her faculty advisor. The student fills out an online approval form, along with the T/D and it is submits it. It is placed in a staging area prior to being uploaded to ProQuest. A verification is created automatically and stored for the availability of the Graduates College and Graduate School of Biomedical Sciences and Professional Studies. If the student desires an embargo, the faculty advisor must approve the embargo ad and embargo request is submitted electronically to the University Archives. The request is emailed to iDEA, so that when the upload of the T/D occurs, it is flagged as unavailable for access during the embargo period. (The ProQuest submission also includes an embargo request.)

When the T/D is sent to the staging area by the student, metadata technicians examine the metadata records in the ProQuest ETD ADM application, which is available on two separate dekstops (in the same physical location within the library). They validate the metadata and allow the acceptance of the T/D to ProQuest and iDEA.

Once the records reside in iDEA, they are ready for creation of a DOI and cataloging information to be created. The metadata manager mints a DOI and cataloging and authority work is performed, creating a MARC record. The record is sent to OCLC Worldcat and the the Ex Libris ALMA catalog. ALMA is a cloud-based library services platform. (Currently the workflow diagram shows the record as being placed in Millenium, but Millenium – the former library catalog – was retired in early August and replaced by ALMA). On a daily basis the ALMA records are loaded via Summon, an interface that mediates between the user and the OPAC. THE OCLC Worldcat, ALMA and loaded ALMA records link to the T/Ds in iDEA.

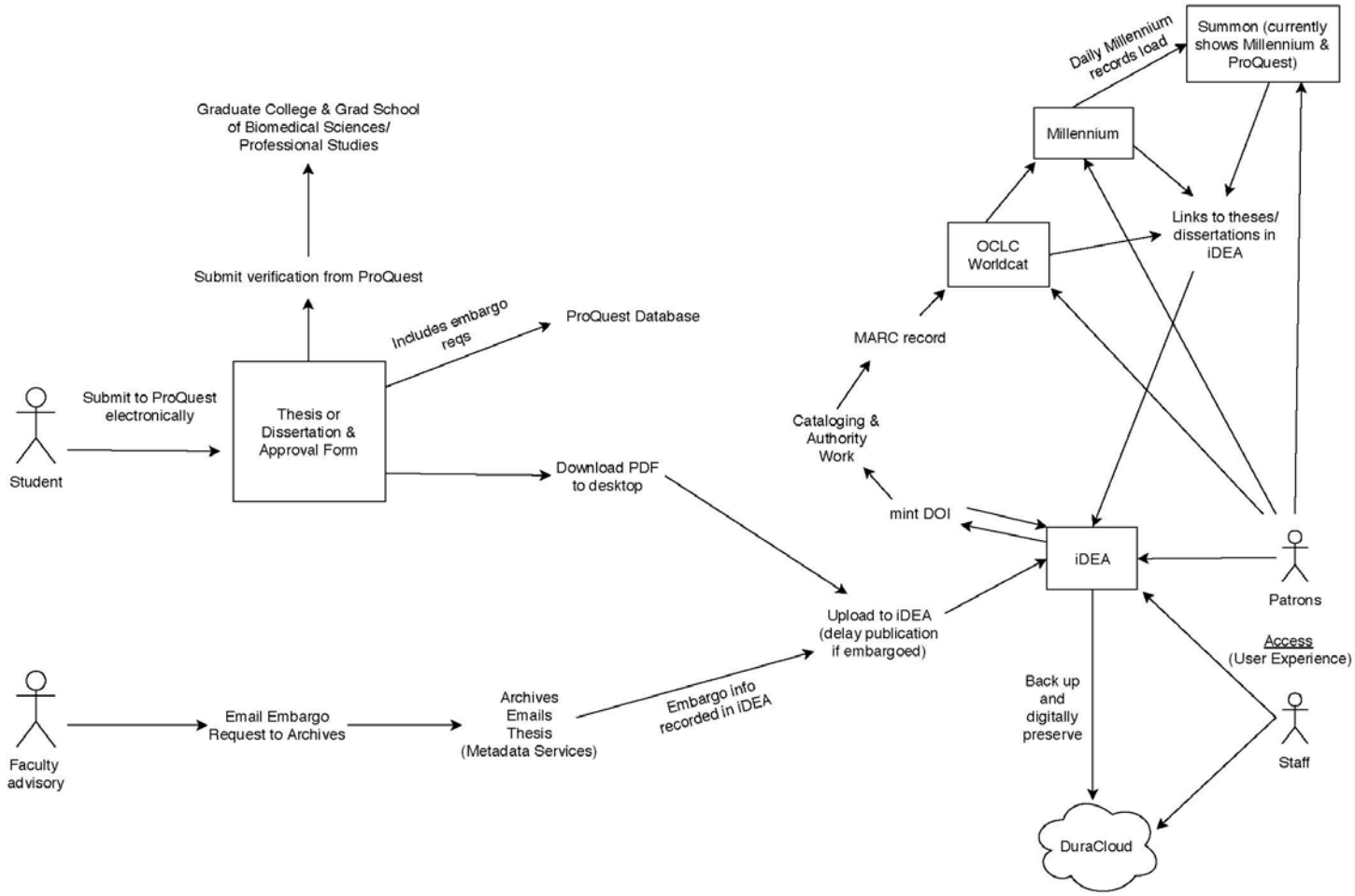

Figure 3 - Drexel University ETD Workflow (Detailed)

iDEA thus provides the storage (the system of record) for the ETDs and the source through which patrons can use Summon to access those records. In fall, 2018, the records in iDEA will be backed up and digitally preserved in DuraCloud, with which the University Libraries already has a contractual agreement. When the preservation environment in DuraCloud is implemented, Islandora will be used for discovery and access. Bundled together are the T/D, its metadata, and a thumbnail.

Currently, the personnel in the CCI Technology Department manage all system administration of the iDEA contents, but the information that resides on their servers will be outsourced to a company called Born Digital, which will also manage the migration. All information residing in the library will be migrated to the Technology Department servers.

Risks and Benefits

Currently, the servers holding the ETDs are all located in a single location, which represents a serious risk to the ETDs. If a catastrophic event were to destroy the environment, all ETDs would be lost. It is recommended that the movement to DuraCloud occur without delay. DuraCloud represents a "true" preservation environment, and offers NDSA or near-NDSA high levels of security.

Beyond this single, serious risk, few risks were found to the workflows and processes that have been developed by the University Libraries. Internal checksums occur when the T/Ds are delivered both to ProQuest and to iDEA, supporting the integrity of the records movement. Two desktops exist with the ETD ADMIN middleware, alleviating the worry of having to replace both the middleware and desktops in the event of disaster. (In fact, this is a very slight risk, since the middleware is readily available and could easily be placed on a new desktop if necessary.) Currently, if an ETD or ETDs are damaged or lost, they must be recovered (at a price) from ProQuest. The cost of recovery is unknown to the author. ProQuest is considered an exemplary service for storage of ETDs and numerous cases exist in which data damage and leakage has been discovered by comparing local records to ProQuest records.

One area of risk that could not be evaluated is that related to the service contracts with the various external service providers. These contracts were unavailable to the author. It is recommended that they be reviewed to ensure consistency with each other and with the policies and procedures located in the Libraries' Digital Preservation policy, developed by Deb Morley and Matthew Lyons. In fact, as mentioned earlier, to begin to come closer to "trustworthy" status, all workflows, policies and procedures should be documented and placed in an environment in which they are readily available for

internal and external personnel. Part of the notion of trust is that of openness, and the processes and policies should be available for any interested user (excepting, of course, some security-related information that must remain confidential to maintain security of the system from hackers.)

Currently, Figure 3 is not entirely accurate, but rather, is a depiction of the future of the ETD workflow. In the lower right hand corner of the workflow diagram, "DuraCloud" is depicted as a resting place for ETDs. In fact, due to resource shortages, that particular portion of the flow is not yet in place. Rather, ETDs find their final resting place in iDEA, the Drexel-developed system. Risk of damage and lost will be reduced significantly when that process is put in place.

Some risk of loss or damage exists currently as a potential due to the possibility of hacking into iDEA. The degree of vulnerability to this possibility is unknown. To mitigate this possibility, regular comparison testing between the ProQuest record and the iDEA record should occur, most efficaciously by selecting a small, random sample of records and checking them, perhaps every six months or by contracting with ProQuest to do a batch comparison on a regular basis, perhaps yearly. CCI Information Technology personnel should ensure that regular comparisons of records occur. Whatever method is chosen to ensure that document integrity and recovery occurs appropriately, it is essential to document the entire process of document recovery.

Some best practices exist that the author was not able to fully examine, and it is recommended that appropriate personnel in the workflow ensure that these best practices are followed. For example:

- Have unique directory names, with timestamps;
- Ensure you use the same naming convention for scanned and born-digital ETDs; (at this point, this is irrelevant for the iDEA system, since conversion of hard-bound T/Ds has not yet occurred);
- Use discreet static (unchanging) archival units (clusters of ETDs) (e.g., annual ingest into preservation caches)
- Suggested accumulation size for archival units of no more than 10 GB for portability"
- Keep ETDs on live, spinning discs

As far as can be seen, these recommendations are in line with current iDEA practice, but this should be doubly verified with internal personnel.

Benefits

The new ETD system will save the Libraries significant storage space in the future. In addition, it will provide wider, easier access to student outputs.

Taking the first step of conducting an NDSA examination of the implemented system was a wise decision. It should be continued on an ongoing basis to ensure that the Libraries continues to move toward more trusted status in the future and to ensure that potential risks which may arise are captured and mitigated quickly.

An additional recommendation is that the Libraries consider moving over time toward the acceptance of less traditional types of student output. More and more theses and dissertations are created that are not "manuscript" in nature. There are numerous multimedia and mixed media dissertations created and this will only grow over time. A process for ensuring the secure storage and access to these types of output will bring Drexel truly into the 21st century way of scholarly thinking as time goes by.

Costs

Some cost estimates were provided by former Data and Digital Stewardship Director Deborah Morley, and this is provided here. ProQuest costs are not included here, because they are a sunk cost – ProQuest has been used as an access (and backup) service even during the time that T/Ds have been stored in hard-copy at the Libraries. Three primary components of costs exist for the ETD program:

- Libraries' staff labor,
- iDEA, the digital repository, and
- DuraCloud Services

Costs for ExLibris ALMA are also excluded here, because it is not used strictly by the ETD program. Further work would be required to estimate how much, if any, of the ALMA costs should be allocated to the service. Likewise, a portion of overhead costs should be included, but is not done so here. To estimate this, a couple of techniques could be done – one could estimate the amount of space used by the program (which would probably yield an insufficient estimate, since most of the costs are due to digital services). Likewise, one could estimate the percentage of librarian and other personnel time as a proportion of overall library time and use that result as the proportion of overhead to represent the ETD service. No perfect way of allocating overhead exists and at best, an estimate can occur.

Libraries' Staff Labor

Current information for costing the Libraries' staff labor is not as granular as it could be in the first estimate shown below. Ideally, it should be calculated on an hourly basis per year. (E.g., although there are 52 months in a year, there is often a two-week vacation allotted to personnel, during which they are not working on their projects.) Thus, it would be reasonable to calculate the salary based upon the average hourly salary times 40 hours per week time actual work hours per year. With a two week vacation this would be 2,000 hours. However, there is usually a loss of time added into this as well, for sick days, time spent away from desk (e.g., "watercooler time, bathroom time, and other unproductive time, etc). Often this implies an actual productivity estimate of between 70% and 80% of the 40 hour work week. The estimates given in the first table use a 40 hour work week. If one calculates by actual hours, the cost figures will decline somewhat. This should be modified if the work week is different than 40 hours.

Estimates used for salaries and time spent by the relevant personnel are

- 10% of one metadata services technician at an average salary of $38,000 per year;
- 25% of another metadata services technician at an average salary of $38,000 per year;
- 2.5% of the metadata librarian's time at an average salary of $64,000 per year;
- 1% of the archives technical at an average salary of $38,000 per year; and
- 2.5% of the university archivist's time at an average salary of $64,000 per year.

The Excel spreadsheet within this document can be changed to provide actual salaries, if desired. Likewise, to calculate Year 2 totals, an inflation factor (for raises) should be included, if desired. Currently, I have estimated a 2% raise.

The second table accounts for productivity loss due to various factors, which are enumerated in the section called "Assumptions."

iDEA

These costs include

- Payments to CCI IT for technical infrastructure and systems administration
- Payments to Born Digital for application support and development of iDEA
- Libraries' staff time who are involved in the management and oversite of iDEA.

The current estimate per year, based on the CCI-Drexel Libraries contract is $38,000, not including the Born Digital costs.

Born Digital

The costs associated with Born Digital are currently $12,000 per year. An inflation factor for Born Digital would likely be estimated at 3%, which is conservative. This information may be available from Born Digital and should be reflected in the Born Digital contract. It is reflected in the Year 2 estimates.

DuraCloud

As with Born Digital, inflation factors are not known, but 3% to 5% would not be surprising. Information regarding fees for DuraCloud came from Deb Morley. Currently, a 3% estimate is given for Year 2. Currently no inflation factor is included for the storage fees, but one should check the contract to determine if those should be included.

Conclusion

This overall risks associated with the ETDs implementation are very low, with the exception of the risk of damage due to a disaster that affects the CCI Technology Department equipment, e.g., fire. Other risks relate to the potential loss of personnel and consequent loss of organizational knowledge, hence the recommendation for trustworthy-level documentation of all policies and procedures. The benefits appear to be high. Storage space is at a premium and, although specific costs of hard-bound storage space are unknown, other organizations have seen significant savings by moving to an electronic environment. Two embedded Excel spreadsheets are included here to see current results. The first shows current costs and projected second year costs, including per volume costs, with the assumption of a 40 hour work week and full productivity. The second shows current costs and projected second year costs, including per volume costs, with the assumption of 79% productivity. The higher per volume costs reflect that productivity reduction.

Library Staff				Year 1	Year 2**			
	Time^^	Average Salary	Benefits (.345)	Total	Total			
Metadata Services Technician	0.1	$ 38,000	$ 13,110	$ 16,910	$ 17,248			
Metadata Services Technician	0.25	$ 38,000	$ 13,110	$ 22,610	$ 23,062			
Metadata Librarian	0.025	$ 64,000	$ 22,080	$ 23,680	$ 24,154		Spring 2018 volumes =	3631
Archives Technician	0.01	$ 38,000	$ 13,110	$ 13,490	$ 13,760		First year per volume cost =	$ 42.14
University Archivist	0.025	$ 64,000	$ 22,080	$ 23,680	$ 24,154			
Total				**$ 100,370**	**$ 102,377**		Second year extrapolation per volume = ^	$ 14.40
iDEA								
CCI IT Charges				$ 38,000	$ 38,000			
Born Digital				$ 12,000	$ 12,360			
Total				**$ 50,000**	**$ 50,360**			
DuraCloud* **								
Annual Subscription				$1,234	$1,234			
DuraCloud Storage subscription*				$1,400	$ 2,885			
Total				**$2,634**	**$ 4,119**			
			Estimated Annual Total	$ 153,004	$ 156,856			

Productivity Factors

When productivity factors are included, total costs go down because less work is assumed to be completed per year on the project. If the total number of volumes remains constant, or grows, this implies that less time will be available for other activities as more time will need to be spent proportionally on the ETD activities to avoid backlog. Simply reducing time effort will reduce the overall costs of the ETD processes, but will increase the cost per manuscript to reflect the reduced effort to maintain the same number of manuscripts processed. This is factored in below. As actual productivity, defined by the number of volumes that can be processed per unit of time increases, overall costs will decline and costs per volume will also decline.

Library Staff				Year 1	Year 2**			
	Time^^	Average Salary	Benefits (.345)	Total	Total			
Metadata Services Technician	0.079	$ 38,000	$ 13,110	$ 16,112	$ 16,434			
Metadata Services Technician	0.1975	$ 38,000	$ 13,110	$ 20,615	$ 21,027			
Metadata Librarian	0.01975	$ 64,000	$ 22,080	$ 23,344	$ 23,811		Spring 2018 volumes =	3631
Archives Technician	0.0079	$ 38,000	$ 13,110	$ 13,410	$ 13,678		First year per volume cost =	$ 49.81
University Archivist	0.01975	$ 64,000	$ 22,080	$ 23,344	$ 23,811			
Total				**$ 96,825**	**$ 98,762**		Second year extrapolation per volume = ^	$ 17.02
iDEA								
CCI IT Charges				$ 38,000	$ 38,000			
Born Digital				$ 12,000	$ 12,360			
Total				**$ 50,000**	**$ 50,360**			
DuraCloud* **								
Annual Subscription				$1,234	$1,234			
DuraCloud Storage subscription*				$1,400	$ 2,885			
Total				**$2,634**	**$ 4,119**			
			Estimated Annual Total	$ 149,459	$ 153,241			

Assumptions Made in this Assessment

1. Including vacation, sick time, personal leave, civic engagement, holidays, and summer and winter breaks as time off, one is left with a 79% FTE productivity. This implies a total work load of 205 days of actual productivity being assumed in the second set of calculations. The first set of calculations assumes a 40 day work week with full productivity.)

Vacation:	20 days
Sick:	12 days
Personal leave:	2 days
Civic engagement	2 days
Holidays:	9 days
Drexel summer break	5 days
Drexel winter break	5 days
Total	55 days

2. DuraCloud Preservation Subscription service. Storage involves 2 TB Amazon S3 @ $700/Terabyte.
3. Second year upgrade to DuraCloud Preservatioin Plus, so that copies are stored on both Amazon S3 and AWS Glacer. Addition cost is $125 per TB per annum.
4. Volumes of submissions are assumed to occur over 3 quarters and to have volumes that approximate each other. (If fall and winter have fewer volumes, this will reduce overall costs.)
5. Wage rate inflation is assumed at 2% per annum.
6. No iDEA inflationary charges are yet included.
7. No DuraCloud inflationary charges are yet included.
8. Actual number of Spring 2018 volumes was 3631 – no assumption, actual number.