Data Desiccation: Facilitating Long-Term Access, Use, and Reuse of ETDs

Mark Edward Phillips
Digital Library Services,
University of North Texas Libraries
Mark.Phillips@unt.edu

Daniel Gelaw Alemneh
Digital Library Services,
University of North Texas Libraries
Daniel.Alemneh@unt.edu

ABSTRACT

The successful management of electronic theses and dissertations (ETDs) requires effort across the entire life-cycle to ensure that ETDs are managed, preserved, and made accessible in a manner that today's users expect. Given the pressure of reading more in less time, today's users demand access to various formats regardless of temporal and spatial restrictions and the types of devices used. Digital curation is the active management of any type of digital resource through its entire life-cycle, from creation and active use, to preservation and re-use. ETDs are a highly specialized collection that demands a more specialized treatment and characterization to better capture the semantic relations of the underlying concepts. Over the past year, the University of North Texas (UNT) Libraries have put forth great effort in making digital collections more accessible and useful in research processes. This paper discusses UNT's ETDs curatorial activities including how ETDs users can benefit from desiccated versions, traditionally discussed only in a digital preservation context.

Keywords

desiccation, long-term access, preservation, ETD

INTRODUCTION

Today, more urgently than ever, researchers need ways to facilitate their research while at the same time promoting scholarly communication within and beyond their own domains. As electronic theses and dissertations continue to be an area of great interest to institutions of higher learning throughout the world, there is an increased need to improve access to these resources across a variety of access systems and tools. The University of North Texas (UNT) Libraries have re-envisioned the level and type of access to the collection of Electronic Theses and Dissertations it houses as part of the UNT Digital Library. UNT has approached this process with the use of a methodology it is calling "data desiccation" as it archives and provides access to the scholarly output of students at UNT. The authors define data desiccation as the systematic migration to formats which provide a limited set of functionality when compared to the original format but which provide a simplified preservation strategy and additional access points to the underlying information. Specific examples are explained later in this paper.

Overview

The University of North Texas (UNT) has a long history with electronic theses and dissertations. Since 1999, the UNT Graduate School has mandated the electronic submission of theses and dissertations as a requirement for graduation from the masters and doctoral programs at the university. Partnering with the Toulouse School of Graduate Studies, the UNT Libraries provide long-term storage and preservation and provisions access to these documents to interested users around the globe.

The Digital Projects Unit of the UNT Libraries began investigating ways to provide additional end user access points to these documents in the summer of 2009 and arrived at the process described in this paper. To date over 4,000 ETDs have been migrated to the new delivery format, and, as new documents are deposited, they receive automatic migration to the new data model.

WORKFLOW

Data desiccation in the context of the UNT ETDs first involves converting the deposited PDF into a series of image files that serve as the primary access point to the documents online. UNT has decided that high quality JPEG images are to be used as the image format for this conversion process. As shown in Table 1, both the TIFF and JPEG2000 file formats were investigated before this decision was made, but each format posed challenges to existing workflows or cost models and did not provide additional preservation or access features.

Format	**Average Image Size**	**Local workflows for processing**
JPEG2000	5 MB	No
TIFF	64 MB	Yes
JPEG	7 MB	Yes

Table 1. File Formats Examined

UNT utilizes Adobe Acrobat Professional (Version 9.1.1) for the conversion of the PDF documents to JPEG. Each image is saved as a 24bit RGB image at 600 dpi; the highest level quality setting (least amount of compression) available in the tool is used when saving the image. All internal color-spaces are converted to RGB for the export. This conversion was carried out on 4,000 files--resulting in more than 500,000 JPEG images. During the automated conversion of files many errors occurred—particularly for the older ETDs created using older versions of Adobe Acrobat software. For those items with conversion errors, UNT employed a number of methods in order to resolve all transformation issues. The most common method used was to print and rescan pages that did not automatically convert during the process. Table 2 depicts some of the common conversion issues together with the possible solutions and corrective measures.

Issue Type	**Error Message**	**Number of Instances**	**Remark**
Pages not able to be converted	"Acrobat Couldn't save a page in this document because of the following errors: Bad Parameter"	164	UNT printed scanned and reprocessed pages. Such error messages occur when saving a file with password security in Adobe Acrobat. UNT solved that by going to Edit> Preferences> Security and unchecking "Verify signatures when the document is opened".
Non-Windows and/or not embedded fonts	"Cannot find or create the font "Times-Bold." Some characters may not display or print correctly."	323	"Times-Bold" is not a font that is typically installed on Windows platforms. This problem seems more prevalent with LaTeX users who would not embed fonts. It is possible that this is a font that is used on Mac or Linux systems. The equivalent font for Windows is "Times New Roman." UNT has corrected the problem with a "preflight" script that handles embedded fonts.
Some other set of type that may not quite affect pdf to jpg conversion	"Cannot find or create the font 'WPMathA'. Some characters may not display or print correctly."	-	Despite these error messages, UNT was able to convert the files. There appears to be a delayed reaction and the error message still displays after the successful completion of pdf to jpg conversion. The only apparent solution to get rid of the annoying error message is to just close and reopen the file. This error does not happen consistently and seems to be more prevalent with LaTeX users.
	"Acrobat failed to send a DDE command"	-	

Table 2. Files with Conversion Issues

Once the image creation has been completed, UNT matches the image sequences to the pagination in the ETD. While somewhat tedious this process reveals a number of issues which have gone unnoticed-- in many cases--for over a decade. For the recording of image sequence and pagination the UNT Libraries use a common file naming convention used throughout the digital library and digital conversion world. Locally this process is known as “magick numbering” of a document. The process involves two running sequences of numbers concatenated into a single filename. Table 3 demonstrates this technique.

Sequence	Pagination	Filename
1	Title Page	000100tp.jpg
2	Copyright page	00020000.jpg
3	Abstract	00030000.jpg
4	ii	000400ii.jpg
5	iii	00050iii.jpg
6	iv	000600iv.jpg
7	1	00070001.jpg
8	2	00080002.jpg
9	3	00090003.jpg
…	…	…

Table 3. Magick Numbering

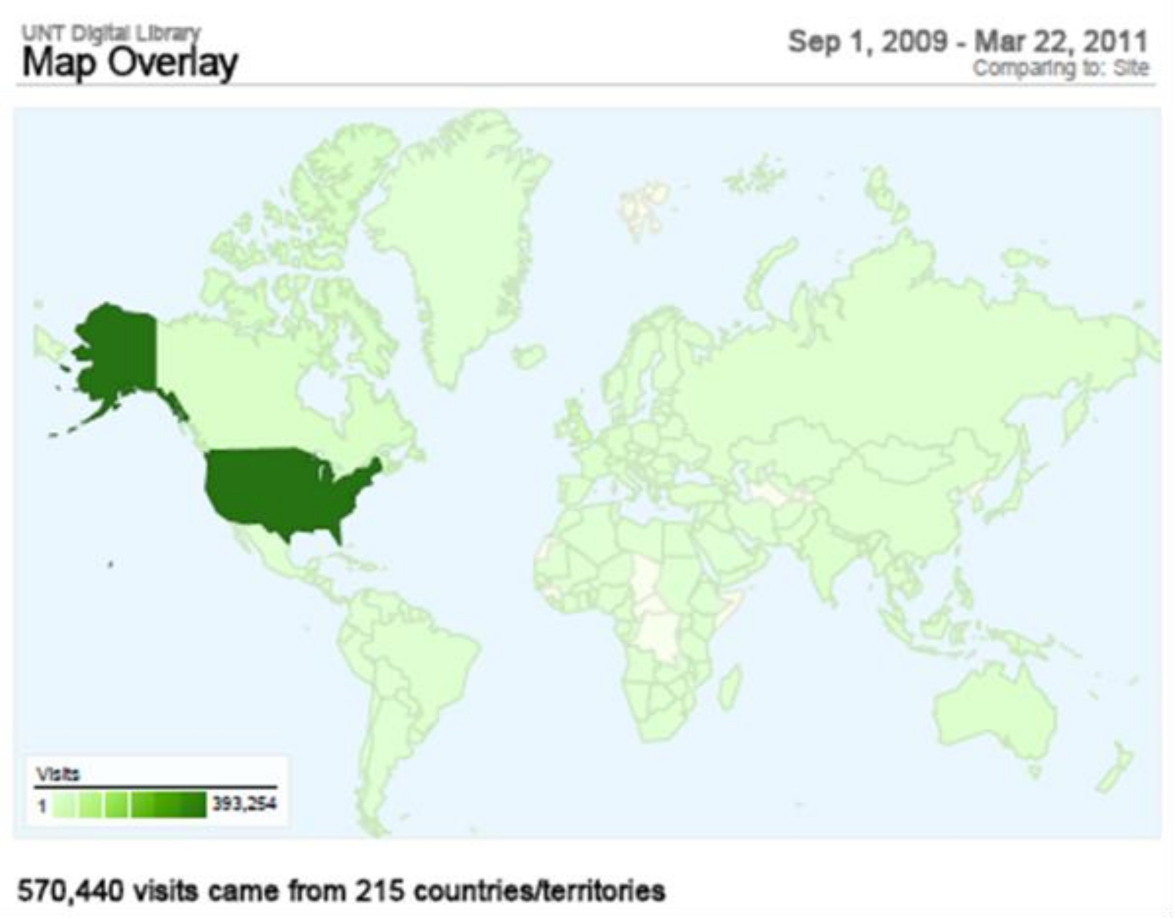
For the ETD conversion an eight digit filename was used, however magick numbers can be easily extended to allow for pagination and sequences above 9,999 pages by moving to a ten digit filename. There are several methods of notating pages numbered with Roman numerals and other document features such as numbered plates, and front, back, and inside. While some of the work for this process can be done in an automated manner, it is still the most time consuming operation in the workflow. One positive side effect of this process is the ease in which mis-numbered or missing pages can be identified--typically a trigger event which may necessitate communication with the Graduate School and possibly an updating of the document held by the Libraries. To date 35 dissertations contained errors that required intervention by the Graduate Scholl in order to move forward with the migration. Once the magick number is completed and any errors have been resolved, the files are run through an optical character recognition (OCR) process utilizing the PrimeOCR engine. The proprietary PRO format from the PrimeOCR engine is then converted into a simple ASCII text file and a UNT-specific word bounding box file. All of these derivative files, in addition to the originally deposited PDF version are combined to form the submission information package (SIP) used for ingest into the repository.

```
alemneh_daniel/
   |-- 01_jpg/
   |   |-- 000100tp.jpg
   |   |-- 000100tp.pro
   |   |-- 000100tp.pro.xml
   |   |-- 000100tp.txt
   |   `-- ...
   |-- 02_pdf/
   |   `-- dissertation.pdf
   `-- metadata.xml
```

When the file is ingested into the UNT Libraries preservation repository infrastructure, Web size derivative files are created such as thumbnail and smaller resolution images, which are presented to the end users of the UNT Digital Library. METS files are created automatically to provide structural information and providing a framework for storing preservation metadata using the PREMIS data dictionary. Tools such as JHOVE and the Unix File utility are used to provide file characteristics, which are additionally stored in the METS file for future use in preservation planning and policy development. The archival information package (AIP) is deposited in the UNT Libraries' digital archive called Coda, and the access content package (ACP) is moved into the UNT-developed Aubrey content delivery system where users around the world access it through the UNT Digital Library interface.

IMPROVING ACCESS

Academic libraries provide services to support the creation, organization, management and use of digital scholarship. Like so many academic libraries, the UNT Libraries are actively and continuously seeking to support research and scholarship at UNT by facilitating and enabling the creation and use of diverse scholarly contents. The UNT Libraries compile system-wide aggregated usage statistics for digital resources they manage. As can be seen from Figure 1, the UNT Digital Library is used more than half a million times by people in over 200 countries around the world. In the UNT Digital Library system, ETDs receive significant usage compared to the overall percentage of digital objects. In the 2010/2011 academic year alone, UNT ETDs were accessed more than 300,000 times from around the world. It is expected that providing JPG format in addition to the original PDF format will facilitate access and further promote the scholarly output of UNT's alumni.



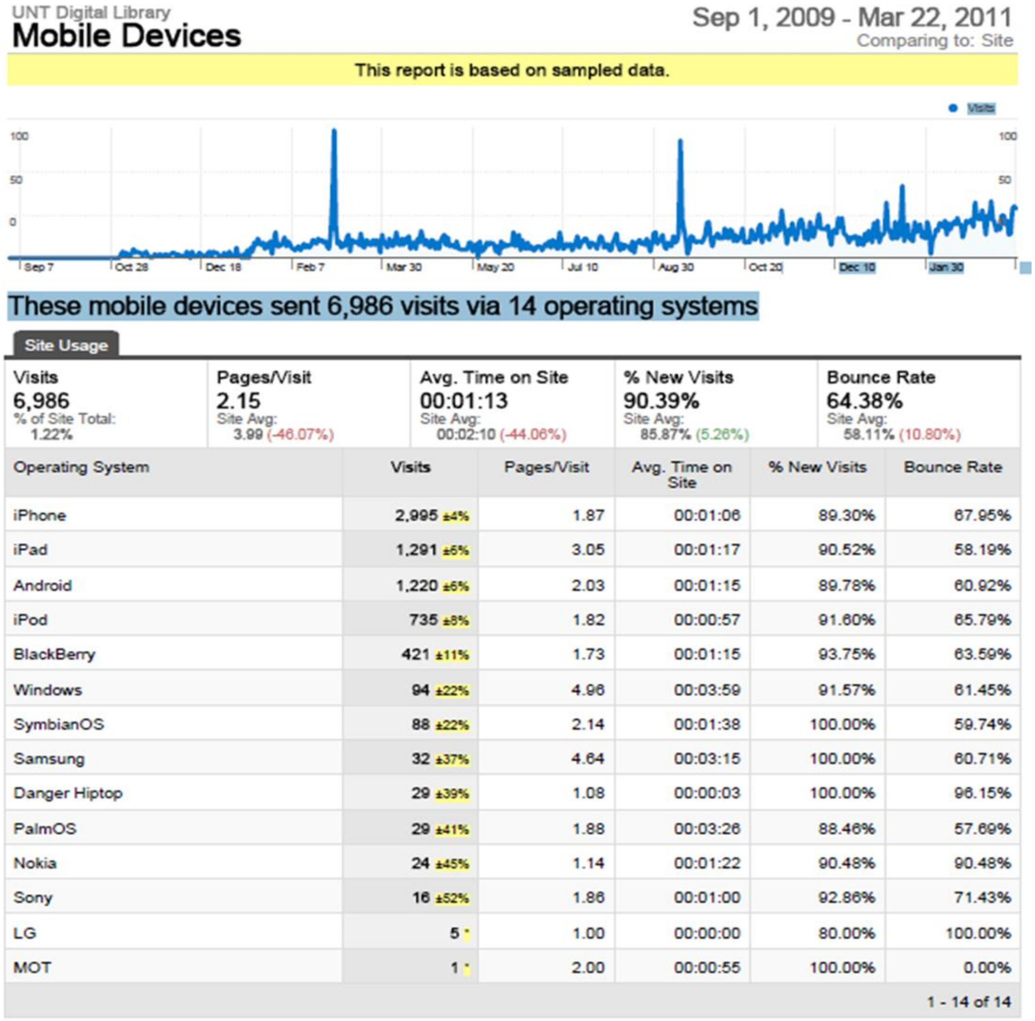
570,440 visits came from 215 countries/territories

Figure 1. UNT Digital Library Access By Countries (from September 2009 to March 2011)

Access Via Mobile Technologies

The proliferation of tablets, mobile phones, connected appliances and other smart machines is driving up the demand for connectivity. Cisco predicts that the number of network-connected devices will be more than 15 billion, twice the world's population, by 2015, (i.e. more than two connections for each person on earth.) It is projected that mobile data traffic will grow 40 times over the next five years and that more people will, access the Web from their mobile devices than from their desktop computers (Cisco 2011).

Before fully adopting, and investing resources in deployment of mobile technology, many institutions try to assess the trends and perform cost benefit analyses. Accordingly, UNT has started preliminary assessments and made Blackboard Mobile available for open testing by faculty and students. Results from a spring 2011 survey of UNT students using Blackboard show that 94.6% of respondents already own smartphone devices, while 2.9% plan to buy one within the next twelve months. A majority of the students also responded that they would definitely use their smartphone devices to complete specific tasks which would perhaps include accessing the digital resources of the UNT Libraries if it were possible to do so. (Summary data from this spring 2011 survey are available here: http://svy.mk/kX8iVD.)



UNT Digital Library
Mobile Devices

Sep 1, 2009 - Mar 22, 2011
Comparing to: Site

This report is based on sampled data.

These mobile devices sent 6,986 visits via 14 operating systems

Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
6,986 % of Site Total: 1.22%	2.15 Site Avg: 3.99 (-46.07%)	00:01:13 Site Avg: 00:02:10 (-44.06%)	90.39% Site Avg: 85.87% (5.26%)	64.38% Site Avg: 58.11% (10.80%)

Operating System	Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
iPhone	2,995 ±4%	1.87	00:01:06	89.30%	67.95%
iPad	1,291 ±6%	3.05	00:01:17	90.52%	58.19%
Android	1,220 ±6%	2.03	00:01:15	89.78%	60.92%
iPod	735 ±8%	1.82	00:00:57	91.60%	65.79%
BlackBerry	421 ±11%	1.73	00:01:15	93.75%	63.59%
Windows	94 ±22%	4.96	00:03:59	91.57%	61.45%
SymbianOS	88 ±22%	2.14	00:01:38	100.00%	59.74%
Samsung	32 ±37%	4.64	00:03:15	100.00%	60.71%
Danger Hiptop	29 ±39%	1.08	00:00:03	100.00%	96.15%
PalmOS	29 ±41%	1.88	00:03:26	88.46%	57.69%
Nokia	24 ±45%	1.14	00:01:22	90.48%	90.48%
Sony	16 ±52%	1.86	00:01:00	92.86%	71.43%
LG	5 *	1.00	00:00:00	80.00%	100.00%
MOT	1 *	2.00	00:00:55	100.00%	0.00%

1 - 14 of 14

Figure 2. Access to UNT Digital Library Via Mobile Devices (from September 2009 to March 2011)

As shown in Figure 2, UNT also collects data about the types of devices users employed to access our UNT Digital Library. Ongoing discussions such as the recent UNT Libraries Tech Talk, “Library on the Go: A Candid Look at the Deployment of

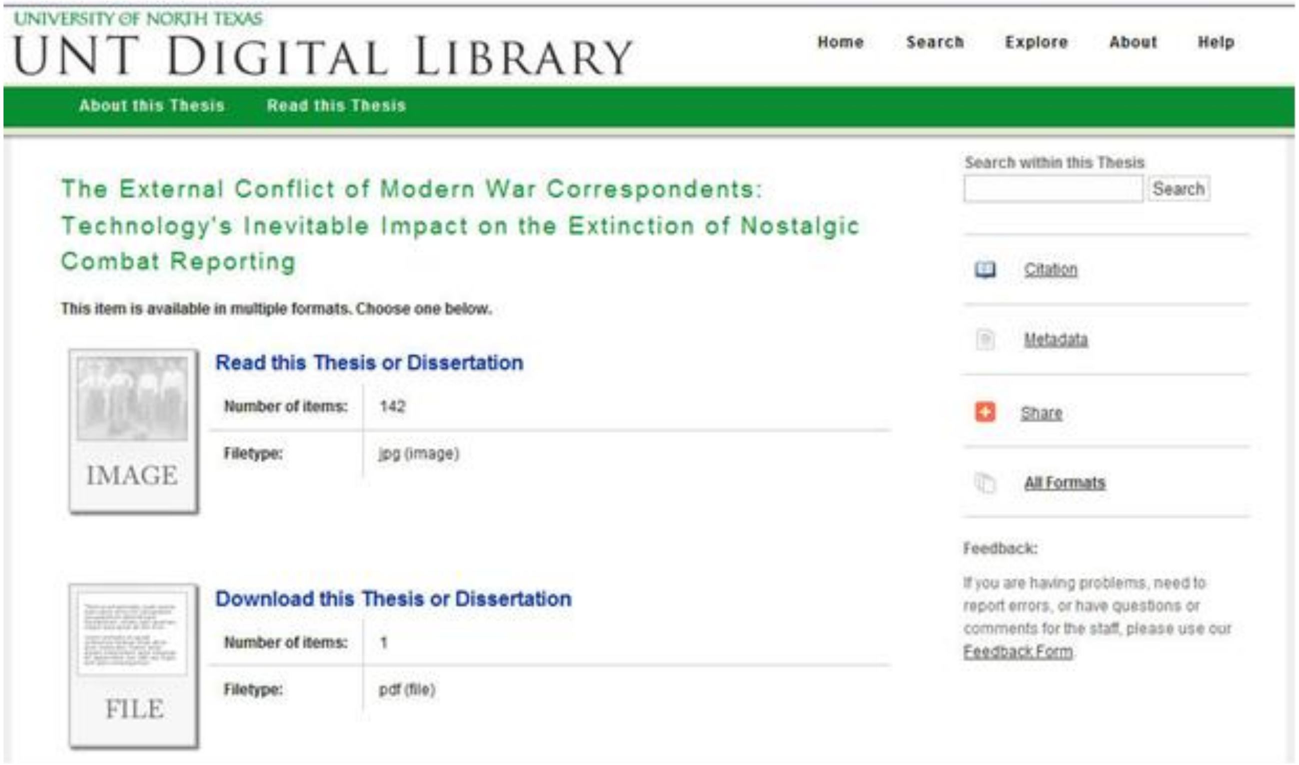
Mobile Technology in the Library" help staff understand users, their mobile devices, the market share of various operating systems, and related trends that influence the use of digital resources.

The multiple format access strategy facilitates access using mobile devices because browsers have the built-in capability to display images, but PDF documents require external applications or plug-ins which may or may not be present. In terms of loading time, the rendering of a PDF document happens only after the complete document is downloaded while images can be streamed in.

Improving Access Via Desiccation

A primary reason UNT decided to engage in this data desiccation process was to provide additional methods of access to the University's ETDs, this improved access manifests itself in two distinct ways. First the Aubrey delivery application exposes the page level OCR text to an increasing number of search engines interested in crawling this content. Second, the user interface provided by the system can take advantage of new and innovative ways of presented paginated book content such as the many image based page turning interfaces such as the GNU Book interface or other interfaces designed for emerging mobile devices such as the iPad or other handheld tables. The user interface in the UNT Digital Libraries has an integrated page turning interface, which provides users with highlighted search terms and other features such as rotation and image magnification. In addition to these new forms of access the originally deposited PDF that is considered to be the "master" format is indexed and available to the end user if that is the preferred method of access.

UNT expects to complete migration of ETDs by spring 2012. Data will be collected to see if desiccated ETDs receive more use than the older, single-format PDF versions did. Considering the synergies of numerous emerging trends, (such as the global development of open access repositories, explosive growth of mobile technologies, cross discipline collaborations, inter-institutional data sharing, etc.), the result of all of this work is expected to be an overall increase in access to the ETDs in the UNT Digital Library.



The External Conflict of Modern War Correspondents: Technology's Inevitable Impact on the Extinction of Nostalgic Combat Reporting

This item is available in multiple formats. Choose one below.

Read this Thesis or Dissertation

Number of items:	142
Filetype:	jpg (image)

Download this Thesis or Dissertation

Number of items:	1
Filetype:	pdf (file)

Search within this Thesis

Feedback:

If you are having problems, need to report errors, or have questions or comments for the staff, please use our Feedback Form.

Figure 3. Facilitating ETDs Access Via Multiple Formats

Improving Long-term Accessibility

In addition to providing new methods of access to the ETDs, the authors believe that the long-term preservation and accessibility of these documents is being improved through the use of data desiccation methods. It is obvious that a JPEG image-based version of a PDF does not provide many of the basic functions available in the master format such as linking, notes, and advanced multi-media capabilities. Despite this fact, the authors believe that simultaneously moving multiple formats into the future will allow the documents to be rendered and access in the future. Providing multiple options certainly, facilitate and promote long-term accessibility. The time consuming process of conversion in the present allows for various format migrations by automated means in the future, and it should be highlighted again that during the conversion process there were many issues that required manual intervention to resolve. This would need to be completed in the future where the likelihood that available software tools will support all features available in the PDF document format is unlikely. In addition to the JPEG version of the document, the OCR text can be used to create a simple text version of the publication even though it is has many levels of functionality removed when compared to the original version. A simple text version of the publication would not provide access to tables, drawings or illustrations but in a hypothetical situation where the other formats were rendered unusable, this would provide at a small level of access to the document.

Future Work

The UNT Libraries feels that the continued use of data desiccation techniques described in this paper are generally applicable in the long-term digital stewardship of content deposited and created by the university. These techniques are also being used in the UNT Scholarly Works collection, which acts as UNT’s institutional repository and adds several new layers of processing from a wider set of deposited file formats. The Digital Library Division is actively seeking new solutions to further streamline the conversion process, there is discussion of a more automated process for matching pagination to the running image sequence and improved conversion tools for creating derivative images from the original PDF files. Finally there is work underway to convert the proprietary PRO format from the PrimeOCR server into the ALTO model used in many projects such as the National Digital Newspaper Program.

CONCLUSION

The innovative use of technologies to provide increased access to and preservation of electronic theses and dissertations attest to the importance of these collections as the scholarly output of institutions of higher education. As an early adopter of what was to become the ETD movement in higher education, UNT encountered and overcame several challenges in the pursuit of providing greater public access to the scholarship conducted at the university. Users from around the world engage with the ETDs created by students at UNT, and as UNT provides additional access points, it expects this trend to increase with time. Understanding user communities, their information needs, and their use behavior will help to move contents into the users’ space and facilitate use of ETDs. The high utility of this collection warrants time devoted to ensuring its preservation and long-term accessibility. Planning for an unknown future in terms of digital preservation can only be made easier by providing multiple options for long term preservation and accessibility to this important set of information.

REFERENCES

1. Bernard J. (2009). Understanding User-Web Interactions via Web Analytics. Synthesis Lectures on Information Concepts, Retrieval, and Services # 6. Retrieved June 08, 2011, from http://dx.doi.org/10.2200/S00191ED1V01Y200904ICR006.
2. Cisco (2011). Global Internet Traffic Projected to Quadruple by 2015. Retrieved June 08, 2011, from http://www.marketwire.com/press-release/global-internet-traffic-projected-to-quadruple-by-2015-nasdaq-csco-1521099.htm
3. Day, M. (2006). The Long-term Preservation of Web Content. In Julien Masanes (Ed.), Web Archiving (pp. 177-199). Berlin Heidelberg: Springer-Verlag.
4. ELPUB (2011). Digital Publishing and Mobile Technologies. 15th International Conference on Electronic Publishing June 22-24, 2011, Istanbul, Turkey. Retrieved June 08, 2011, from http://www.elpub.net/
5. Internet Archive (2011). Internet Archive Book Reader. Retrieved June 10, 2011, from http://openlibrary.org/dev/docs/bookreader
6. UNT (2011), Student Preferences for Accessing Blackboard via Smartphone: Survey Result. Retrieved June 08, 2011, from http://svy.mk/kX8iVD.
7. UNT Libraries' Metadata Projects Documentation (2011). Retrieved June 08, 2011, from http://www.library.unt.edu/digitalprojects/metadata
8. UNT Electronic Theses and Dissertations (ETDs). Retrieved June 08, 2011, from http://digital.library.unt.edu/explore/collections/UNTETD/browse/
9. Weng, N. (2011). Library on the Go: A Candid Look at the Deployment of Mobile Technology in the Library. Techtalk Presentation at the UNT Libraries. Retrieved June 08, 2011, from http://www.library.unt.edu/digitalprojects/tech-talks/library-on-the-go/