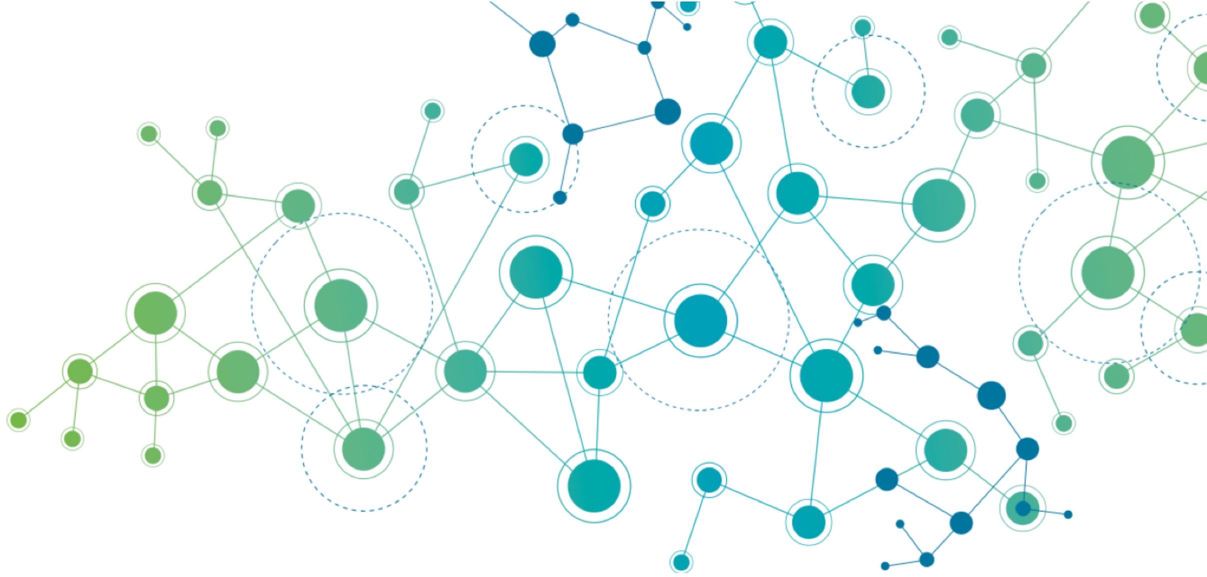


Guidance Documents for Lifecycle Management of ETDs

Authors: Daniel Alemneh, Bill Donovan, Martin Halbert, Yan Han, Geneva Henry, Patricia Hswe, Gail McMillan, Xiaocan (Lucy) Wang

Editors: Matt Schultz, Nick Krabbenhoeft, and Katherine Skinner

18 March 2014

Version 1.0

Publication Notes

Title: Guidance Documents for Lifecycle Management of ETDs

Editors: Matt Schultz, Nick Krabbenhoeft, and Katherine Skinner

Authors: Daniel Alemneh, Bill Donovan, Martin Halbert, Yan Han, Geneva Henry, Patricia Hswe, Gail McMillan, Xiaocan (Lucy) Wang

Publisher: Educopia Institute, 1230 Peachtree Street, Suite 1900, Atlanta, GA 30309.

Table of Contents

Introduction

Matt Schultz and Katherine Skinner (Educopia Institute)

About The Guidance Documents

Over the last fifteen years, colleges and universities have been transitioning from physical (paper/microfilm) to digital submission and management processes for student theses and dissertations. Increasingly, they are accepting and archiving *only* electronic versions of their students' theses and dissertations. While this move from print-based to digital-based theses and dissertations greatly enhances the accessibility and sharing of graduate student research, it also raises grave concerns about the potential ephemerality of these digital resources. How will institutions ensure that the electronic theses and dissertations they acquire from students today will be available to future researchers?

In 2011, a research team led by the University of North Texas, the Educopia Institute/MetaArchive Cooperative, and the worldwide Networked Digital Library of Theses and Dissertations (NDLTD), began studying the production, dissemination, and preservation of Electronic Theses and Dissertations (ETDs). Our original intent was to develop and disseminate documentation for academic libraries that would help curators better understand and address the preservation challenges presented by these new digital collections.

As researchers from the libraries of University of North Texas, Virginia Tech, Rice University, Boston College, Indiana State University, Penn State, and the University of Arizona began to grapple with ETD lifecycle management issues, they quickly realized that librarians were but one of many academic stakeholder groups that work collaboratively to produce and maintain ETD collections. Studying the library role in isolation was neither feasible nor helpful. The scope of our work increased to encompass the roles and responsibilities of core stakeholders in the ETD lifecycle: students, faculty, administrators, technologists, commercial vendors, and librarians.

The resulting *Guidance Documents* address areas of interest to ETD program planners, managers, and curators. They will help this extended set of stakeholders understand, document, and address the administrative, legal, and technical challenges presented by ETDs – from submission to long-term preservation.

We greatly appreciate the Institute of Museum and Library Services' generous support of this project. It is our hope that readers find the Guidance Documents useful in their local work to build and refine their ETD programs.

This *Introduction* to the *Guidance Documents* provides a brief description of each *Guidance Document*. To help different stakeholders target sections of specific interest within them, we have included a *Roadmap*. We also include a section defining the key terms of interest in this publication, *Defining "ETDs" and "Lifecycle Management"*

Chapter 1: Guidelines for Implementing ETD Programs – Roles and Responsibilities

Xiaocan (Lucy) Wang of Indiana State University provides a broad, detailed summary of the types of stakeholders that are involved in the formation and maintenance of ETD Programs and then describes the functions each of these stakeholders might play in key phases of ETD lifecycle management.

Chapter 2: Guide to Access Levels and Embargoes of ETDs

Geneva Henry of Rice University offers a comprehensive study of policies and practices related to access levels and embargoes of ETDs. Henry documents the rationale behind access restrictions (and arguments against them), compares implementations of embargoes/restrictions across different institutions, and considers the roles of different stakeholders in determining how to establish and maintain access restrictions.

Chapter 3: Briefing on Copyright and Fair Use Issues in ETDs

Patricia Hswe of Penn State considers the impact of copyright and fair use on the submission, dissemination, and preservation of ETDs, including the responsibilities colleges and universities have to provide students with clear guidance on their own intellectual property rights. This briefing describes copyright and fair use issues from the student-author's perspective and advises on training/education responsibilities within the academic institution. It also considers the copyright issues that may arise in working with vendors (e.g., ProQuest).

Chapter 4: Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs

Yan Han of University of Arizona provides a comprehensive overview of evaluation practices for ETD collections and articulates the value of collecting and using metrics to establish the value of ETD programs. Han describes quantitative and qualitative approaches that institutions might consider to help assess user behavior and content delivery success for ETD collections.

Chapter 5: Managing the Lifecycle of ETDs: Curatorial Decisions and Practices

Bill Donovan of Boston College describes selection principles, risk factors, and policy decisions that institutions make in order to strengthen the long-term outlook for their ETD collections. Covering a diverse range of curatorial topics including file formats, content organization, migration, normalization, and management of complex (multi-file) content objects, Donovan provides a snapshot of the curatorial decisions that librarians working with ETDs must understand in order to provide strong lifecycle management services to their campus.

Chapter 6: Metadata for ETD Lifecycle Management

Daniel Alemneh of the University of North Texas describes how "metadata," or cataloging information about files, are used in the lifecycle management process. Alemneh provides an overview of ETD metadata practices, discusses what metadata elements are most important in lifecycle management, and documents different stakeholder roles and responsibilities in the creation and maintenance of this information.

Chapter 7: Guide to ETD Program Planning and Cost Estimation

Gail McMillan of Virginia Tech elaborates on the crucial role that economics plays in the establishment, maintenance, and ongoing justification of an ETD Program. McMillan identifies the cost categories associated with ETD lifecycle management, focusing especially upon personnel and technical expenses. The guide different ETD implementation channels, including repository software options and internal/external hosting arrangements, and considers the cost and value associated with each. Finally, McMillan provides case studies based on five institutions.

Chapter 8: Guide to Options for ETD Programs

Dr. Martin Halbert of the University of North Texas documents the spectrum of ETD program implementation and offers guidance for academic decision-makers who are either creating or modifying ETD programs. Dr. Halbert identifies and offers in-depth analysis regarding the five key decisions that ETD programs must make. He also provides a literature review of publications, standards and reports that have been produced to date, and relates these to the key decisions.

Roadmap

The authors have aimed to be comprehensive in their treatment of ETD programs, and we encourage readers to review all of the *Guidance Documents* to gain a holistic view. However, we have also highlighted the sections of each document relevant to four roles in ETD programs:

Administrators

Institutional administrators, deans, associate deans, and other high-level staff responsible for management and oversight

Topic	Section Numbers (Beginning Pages)
ETD program stakeholders and the planning process	*1.1* (1-1), *1.2* (1-2), *1.3* (1-6)
Reasons for and against access restrictions	*2.2* (2-2), *2.3* (2-10)
Intellectual property rights for authors and institution	*3.2* (3-2)
Benefits of program usage statistics	*4.1* (4-1), *4.5* (4-12)
Long-term risks to accessibility	*5.1* (5-1), *5.2* (5-2)
Metadata in the ETD lifecycle	*6.1* (6-1)
Personnel and technical costs in ETD programs	*7.2* (7-3), *7.3* (7-5)
Important decisions in planning an ETD program	*8* (8-1)

Submission Staff

Graduate school and library staff responsible for interfacing directly with authors during ETD creation and submission

Topic	Section Numbers (Beginning Pages)
Other stakeholders and submission responsibilities	*1.2* (1-2), *1.3* (1-6)
Access restriction policy guidance	*2.2* (2-2), *2.3* (2-10), *2.4* (2-11)
Intellectual property rights issues for authors	*3.2* (3-2)
How usage statistics support access policies	*4.1* (4-1), *4.5* (4-12), *4.6* (4-14)
How format policies affect long-term access	*5.1* (5-1), *5.3* (5-4)
How information about ETDs is recorded	*6.3* (6-6)
Personnel costs and program case studies	*7.2* (7-3), *7.4* (7-9)

Access & Repository Staff

Graduate school and library responsible for managing the long-term access and storage of ETDs

Topic	Section Numbers (Beginning Pages)
Other stakeholders and submission responsibilities	*1.2* (1-2), *1.3* (1-6)
Processes for restricting and releasing access	*2.4* (2-11), *2.5* (2-12), *2.6* (2-13)
Intellectual property rights in relation to IRs	*3.2.7* (3-10), *3.2.8* (3-11)
Purpose and methods to collect usage data	*4.1* (4-1), *4.2* (4-2), *4.3* (4-10), *4.4* (4-10), *4.5* (4-12)
Long-term access risks and mitigation strategies	*5.1* (5-1), *5.2* (5-2), *5.3* (5-4), *5.4* (5-9), *5.5* (5-12), *5.6* (5-14)
Metadata standards and workflows for creating metadata	*6.1* (6-1), *6.2* (6-1), *6.3* (6-6), *6.4* (6-10), *6.5* (6-16), *6.6* (6-18)

IT Staff

Graduate school, library, and IT department staff responsible for the technical infrastructure of the ETD program

Topic	Section Numbers (Beginning Pages)
Other stakeholders while planning the program	*1.2* (1-2)
Methods for automated capture of usage statistics	*4.2* (4-2), *4.3* (4-10)
Risks during data migration scenarios	*5.5* (5-12)
Protocol for federating ETD metadata	*6.4.1.5* (6-15)
Costs for IT infrastructure	*7.3* (7-5)

Defining "ETDs" and "Lifecycle Management"

Theses and dissertations comprise an essential record of the intellectual output of students and the mentorship provided by faculty to students in a college or university setting. In the US context, theses and dissertations include three main types of scholarly content – undergraduate honors theses, masters theses, and doctoral dissertations.[1] They are submitted by students in support of their candidacies for academic degrees and to demonstrate their professional qualifications as graduates of an institution.

The term "Electronic Theses and Dissertations" is used commonly in academia to denote a *digital* collection of these formal documents. These "ETDs" may be digitized or born digital; indeed, most academic institutions manage (or will manage) both forms. The use of the term "ETD" primarily differentiates between analog theses and dissertations (paper, microfilm) and their digital counterparts (digital objects).

On the surface, this seems like a simple shift in format, particularly given that to date, the intellectual content of a "thesis" or "dissertation" is fundamentally unchanged – most institutions continue to support a heavily text-based submission that conforms to long-held standards and print conventions.[2] However, in reality, this shift already presents a number of challenges and requires attention to a wide range of legal, administrative, and technical issues.

When colleges and universities decide to support ETDs, most begin by implementing an "ETD Program" involving multiple stakeholders (including the College/Graduate School and the Library) to ensure consistency in the submission, dissemination, and long-term management of ETDs. These local programs provide policies, workflows, and services around such crucial functions as deposit, documenting approvals, metadata capture, rights management, and ingest into commercial and/or library-based repository for management. These ETD programs ultimately are geared toward "lifecycle management" tasks.

Information "lifecycle management" has become an important concept (or set of concepts) that help curators focus their activities and properly assign resources to ensure that information remains accessible and usable over time. Lifecycle management models study and document the progression of digital objects through stages of creation, dissemination, use, update and re-use, storage retention or archiving, and sometimes destruction or disposal, of digital objects.

Some lifecycle management models present themselves as being simple, straight-forward and linear with fairly discrete phases of activity (e.g., Federal Law 44 U.S.C. 2901 and ISO 15489, see *Figure 1*). Other models are more cyclical in nature with overlapping phases depicted (DCC Curation Lifecycle, see *Figure 2*).

[1] In other countries, the terms "thesis" and "dissertation" are often reversed, with "theses" representing the work completed for a PhD.

[2] See Lippincott and Lynch for discussion of the relative inertia of the thesis/dissertation as an academic form: "ETDs and Graduate Education: Programs and Prospects." *Research Library Issues* 270 (June 2010). http://publications.arl.org/rli270/7.

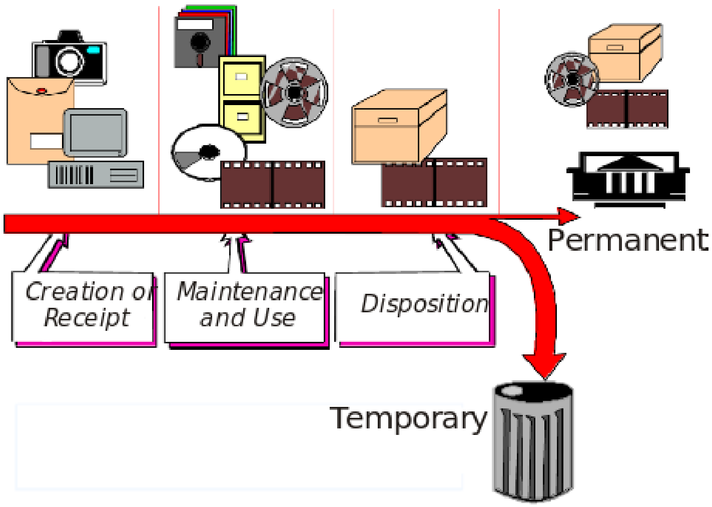


Figure 1. Diagram of the Federal Enterprise Architecture Records Life Cycle[3]

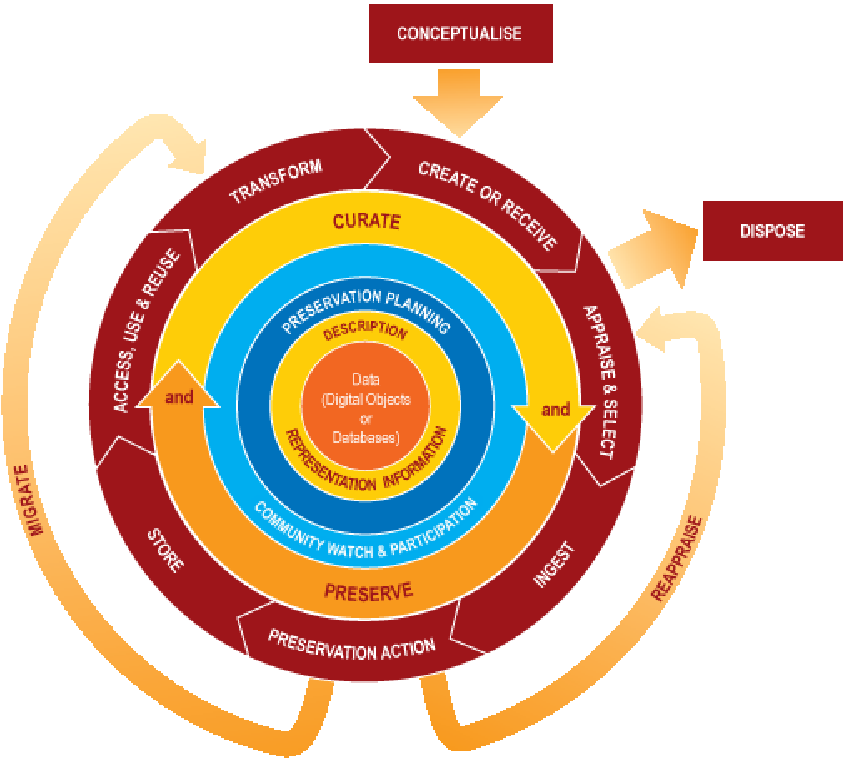


Figure 2. Diagram of the DCC Curation Lifecycle Model[4]

[3] National Archives and Records Administration, Office of Management and Budget, and Federal Chief Information Officers Council, "Federal Enterprise Architecture Records Management Profile." (December 2005). http://www.archives.gov/records-mgmt/pdf/rm-profile.pdf.

[4] Digital Curation Centre. "DCC Curation Lifecycle Model." http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Most models acknowledge that processes, particularly with respect to electronic documents, do not always occur in sequence and that multiple processes can sometimes occur simultaneously or in different orders.

Many stakeholders participate in the lifecycle management of ETDs:

- Student authors *create and submit* ETDs with software applications according to policies (e.g., what formats are allowed).
- Faculty members supervise a student's ETD, mentor students about both the ETD process and scholarly communication in their discipline, serve on dissertation/thesis committees, and participate in policy decisions regarding ETDs at the department, college, faculty, senate, and institutional levels.
- Graduate schools *process, approve*, *embargo, release,* and *update* ETDs over time via online submission systems.
- Libraries/IT/Vendors *catalog, archive,* and *disseminate* ETDs through institutional repository systems and preservation policies/systems.
- Scholars and researchers *use* and *re-use* ETDs via web browsers, download applications, and analysis tools.

In these *Guidance Documents*, we use the phrase "lifecycle management of digital data" in the broad sense defined by the Library of Congress to refer to the "progressive technology and workflow requirements needed to ensure long-term sustainability of and accessibility to digital objects and/or metadata" (Library of Congress 2006). Lifecycle management in this sense is about actively stewarding, through policies, staffing, resources and technologies, a set of digital resources over time.

In the *Guidance Documents* that follow, we address areas within this "lifecycle" that are of special interest, as identified by ETD program planners, managers, and stakeholders.[5] The documents will provide both a series of non-prescriptive strategies that ETD curators can adapt for their ETD programs, as well as pointers to real world examples and demonstrable resources.

[5] McMillan, 2008; Skinner and McMillan, 2009.

1 Guidelines for Implementing ETD Programs – Roles and Responsibilities

Xiaocan (Lucy) Wang (Indiana State University)

Topics Covered

- Potential internal and external stakeholders of an ETD program.
- Reasons and methods to advocate for the establishment of an ETD program.
- Stakeholders to consult in establishing policies and workflow for ETD submissions and ingestion.
- Methods to promote and enhance access to ETD collections.
- Concerns and methods in maintaining long-term access to ETD collections.
- Metrics by which each stakeholder can evaluate their portion of an ETD program.

1.1 Introduction

Since the mid-1990s institutions have increasingly required students to submit theses and dissertations in electronic format(s). The management of electronic theses and dissertations (ETDs) raises a number of issues concerning the processes of ETD creation, ingestion, access, archiving and preservation. As Joan Lippincott (CNI) has noted, institutions that implement an ETD program must carefully consider how best "[a]n ETD program provides a process(es), standards, and software to automate functions, as well as a digital infrastructure for access and preservation" (Lippincott 2006). Like any other project, implementing an ETD program requires the identification of various stakeholders, who have an interest in the success of the program; and the specification of each player's role and responsibility throughout the lifecycle of ETD management. Effectively engaging stakeholders in project management, and successfully coordinating participants' roles and responsibilities, are the keys that enable an ETD program to thrive over time. Without these crucial components, an ETD program can fail at the initial planning stage or lack continued support for further development. Maybe more alarming and more prevalent than either of these fates is that of an ETD program just hobbling along and not meeting the needs of students and researchers and its institution because of poor implementation.

This document provides guidance for identifying potential stakeholders and for understanding their functions at different ETD management phases. It is hoped that the document will be useful for institutions that are beginning to think about an ETD program or just initiating the planning process. Institutions that have implemented an ETD program can use these documented roles and responsibilities to examine their ETD programs and perhaps make some positive modification to their current practices. It is understood that the governance, organization, staffing, policies, and terms differ from institution to institution and from country to country; the involved parties and their functions in a particular locale may not be identical to those specified in the document.

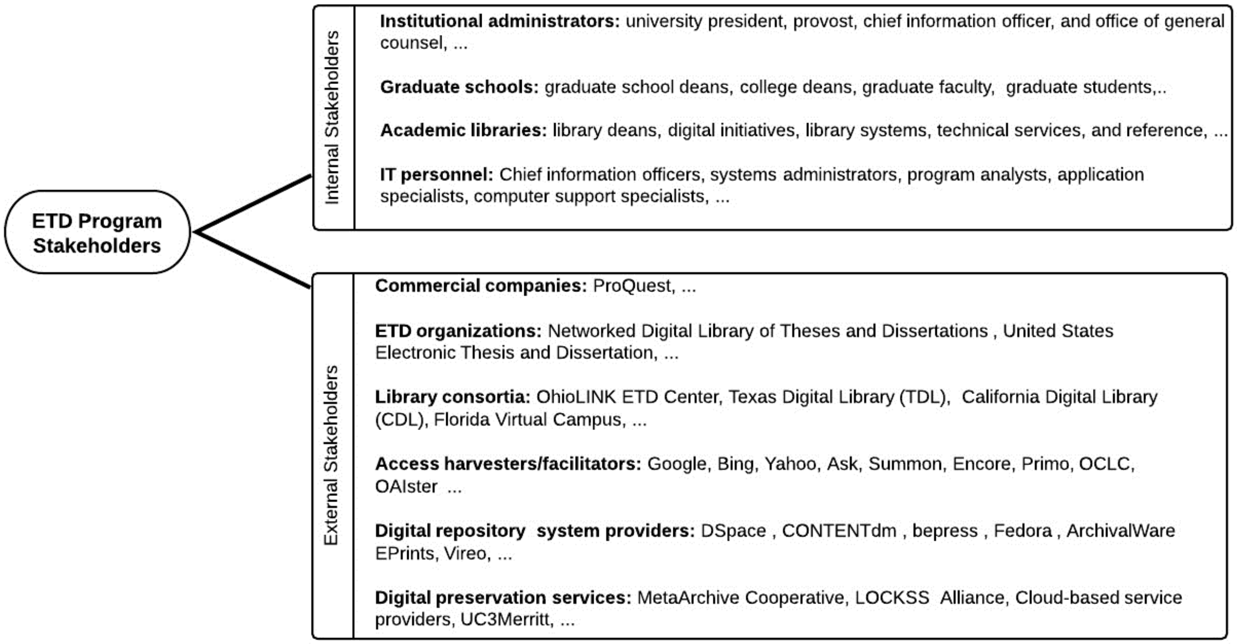
1.2 Types of Stakeholders

Different types of stakeholders have different interests and concerns in an ETD program. Some parties may be actively involved throughout the entire lifecycle of the program, while others may take part in one or two particular processes, either directly or indirectly. Based on whether the stakeholders are from the institution where ETDs are generated, the stakeholders may be broadly divided into two groups: internal and external stakeholders. (see *Figure 1-1* for a diagram of these groups)

Internal stakeholders are the individuals or academic units from institutions of higher education where ETDs are generated. The primary internal stakeholders consist of institutional administrators, graduate schools, libraries, and IT personnel.

Institutional administrators are a group of top-level decision makers such as the university president, provost, chief information officer, and representatives from graduate council and the office of general counsel. Institutional administration personnel are not involved in the day-to-day operation of an ETD program. Rather, they support the program in various ways, including provide general oversight and/or funding support. They may also be the links that ensure the cooperation among other stakeholders.

a. **Graduate schools** are stakeholders directly engaged in ETD programs, especially in the process of planning, creation, and submission. Besides the graduate council previously mentioned, this group includes graduate school deans, assistant and/or associate deans, deans from various colleges or schools, and graduate school staff who handle many details surrounding ETD programs (for example, student service officers and graduate research assistants). In addition, this group includes two other important stakeholders: graduate students and graduate faculty, both of whom are intimately involved in the development of theses and dissertations.
b. **Academic libraries** have been one of the implementers of ETD programs in higher education institutions. Library administrators (i.e., library deans/directors, assistant and/or associate deans, department heads) and departments such as digital initiatives, systems, technical services, and reference together play an important role in ETD advocacy, ingestion, access, preservation, and assessment. Due to internal structures and resource availability, libraries may not have the exact configuration or personnel mentioned above. Likely, academic libraries are group that coordinates or tracks all the responsibility areas related to an ETD program.
c. **IT personnel** also have a stake in ETD programs. Chief information officers, systems administrators, program analysts, application specialists, computer support specialists as well as IT help desk staff are vital to implementing ETD programs. IT personnel may be in a centralized university unit and/or the library's IT unit. Management of born-digital student research papers and retrospectively digitized theses and dissertations demand strong technical support from these information professionals as ETD–related activities require running software applications and server hardware in a network environment.



Institutional administrators: university president, provost, chief information officer, and office of general counsel, ...

Graduate schools: graduate school deans, college deans, graduate faculty, graduate students,..

Academic libraries: library deans, digital initiatives, library systems, technical services, and reference, ...

IT personnel: Chief information officers, systems administrators, program analysts, application specialists, computer support specialists, ...

Commercial companies: ProQuest, ...

ETD organizations: Networked Digital Library of Theses and Dissertations , United States Electronic Thesis and Dissertation, ...

Library consortia: OhioLINK ETD Center, Texas Digital Library (TDL), California Digital Library (CDL), Florida Virtual Campus, ...

Access harvesters/facilitators: Google, Bing, Yahoo, Ask, Summon, Encore, Primo, OCLC, OAIster ...

Digital repository system providers: DSpace , CONTENTdm , bepress , Fedora , ArchivalWare EPrints, Vireo, ...

Digital preservation services: MetaArchive Cooperative, LOCKSS Alliance, Cloud-based service providers, UC3Merritt, ...

Figure 1-1. ETD Program Stakeholders

External stakeholders are entities involved in ETD programs at various levels of engagement. They reside outside the students' institution. External stakeholders exist in various forms: organizations, industrial firms, associations, and individuals. They may be for-profit or not-for-profit. External stakeholders are categorized as follows:

a. **Commercial companies** have vested interests in the publication of ETDs. A leading organization in this group is ProQuest (formerly called University Microfilms International), which has been in the business of centrally collecting dissertation research and distributing microfilm and print copies of dissertations since its founding in 1938 (Lippincott and Lynch 2010). In addition, there are several other commercial enterprises that publish ETDs. One of them is Dissertation.com, a Florida-based company founded in 1997.[1]

b. **ETD organizations**

    a. **Networked Digital Library of Theses and Dissertations (NDLTD)** was established as a voluntary international organization in 1996. Its mission is to promote the adoption, creation, use, dissemination, and preservation of ETDs as well as to support the development of ETD programs (Networked Digital Library of Theses and Dissertations).[2]

[1] Dissertation.com has published more than 600 master's theses and doctoral dissertations as of September 2012, in association with Amazon.com, http://www.dissertation.com/browse.php (last accessed 11-14-2012).

[2] As of April 2012 the NDLTD had reached ninety institutional members, three consortial members and twenty-four individual members from all over the world, including seventy-one universities and institutions. Approximately 80 percent of the institutional members are based in the United States.

b. **United States Electronic Thesis and Dissertation Association (USETDA)**[3] is a non-profit association, established in 2009. One of its missions is to enable and encourage state-wide ETD associations, for example, California Electronic Theses and Dissertations (CAETD),[4] Florida Electronic Theses & Dissertation Association (FLETDA),[5] (Ohio Electronic Theses and Dissertations Association (OETDA),[6] and Texas ETD Association (TxETDA[7]), and to promote ETD program information sharing and advancement.

c. **Library consortia** support local or regional ETD programs, usually by providing ETD submission systems, delivering federated ETD searching and retrieval, and preserving ETDs in a collaborative and cost-effective manner. Some examples are the OhioLINK ETD Center,[8] the Texas Digital Library (TDL),[9] the California Digital Library (CDL),[10] and the Florida Virtual Campus.[11]

d. **Access harvesters/facilitators** are involved in ETD initiatives with emphases on promoting ETD readership and facilitating the processes of searching ETD literature. Access harvesters include major search engines such as Google, Bing, Yahoo, and Ask. Access facilitators comprise web discovery tool service providers (e.g., Serial Solutions' Summon, Innovative Interface's Encore, MANGO, and Ex Libris Primo). Two other stakeholders are OCLC and the OAIster harvesting group that utilizes the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to aggregate ETD metadata from multiple ETD archives.

e. **Digital repository system providers** generally provide a platform for ETD management, including functions for ETD submission, ingestion, dissemination and retrieval. Example software solutions developed by these providers are DSpace,[12] CONTENTdm,[13] bepress,[14] Fedora,[15] ArchivalWare,[16] EPrints,[17]and Vireo,[18] some of which are open source and others of which are proprietary.

f. **Digital preservation services** directly or indirectly archive and preserve digital collections to ensure continued access to digital materials as long as necessary (Beagrie and Jones 2002). The

[3] See http://www.usetda.org/ (last accessed 01-25-2013).
[4] California Electronic Theses and Dissertations, see https://sites.google.com/site/caetds/ (last accessed 01-24-2013).
[5] Florida Electronic Theses & Dissertation Association, see http://www.fletda.org/ (last accessed 01-25-2013).
[6] Ohio Electronic Thesis and Dissertation Association, see http://www.oetda.org/ (last accessed 01-25-2013).
[7] Texas ETD Association, see http://txetda.wordpress.com/ (last accessed 01-25-2013).
[8] The OhioLINK ETD Center was launched in 2001 as a joint project of OhioLINK and the Regents Advisory Committee on Graduate Study, http://etd.ohiolink.edu/faq.html#what-is (last accessed 11-15-2012).
[9] The Texas Digital Library, a consortium of 15 higher education institutions in Texas founded in 2005, http://www.tdl.org/members/ (last accessed 11-16-2012).
[10] See California Digital Library, http://www.cdlib.org/ (last accessed 11-21-2012).
[11] See Florida Virtual Campus, http://fclaweb.fcla.edu/ (last accessed 11-21-2012).
[12] See http://www.dspace.org/ (last accessed 01-25-2013).
[13] See http://CONTENTdm.org/ (last accessed 01-25-2013).
[14] See http://www.bepress.com/ (last accessed 01-25-2013).
[15] See http://fedora-commons.org/ (last accessed 01-25-2013).
[16] See http://www.archivalware.net/ (last accessed 01-25-2013).
[17] See http://www.eprints.org/us/ (last accessed 01-25-2013).
[18] See http://tdl.org/etds/ (last accessed 01-25-2013).

following stakeholders have taken part in the digital preservation management of ETD collections:

a. **MetaArchive Cooperative** is an international membership association founded in 2004[19] that is dedicated to preserving a broad range of digital assets including ETDs.[20] As of December 2012, it serves more than 50 institutional members in 13 states and four countries.[21]

b. **LOCKSS[22] Alliance**, based at Stanford University Libraries, is an international community initiative, committed to providing digital preservation tools and support for digital materials such as ETDs via Private LOCKSS Networks[23] among its members (Stanford University Libraries).

c. **Cloud-based service providers** have stepped into the digital preservation arena with some preservation functionality. Examples of providers include Amazon and DuraCloud.[24] DuraCloud was launched in 2011 and is currently in use by a number of major institutions such as MIT for digital preservation and access to digital scholarship, including ETDs, in a broad range of formats.[25]

g. **UC3Merritt**,[26] developed by the University of California Curation Center, provides long-term preservation of digital assets. The Merritt preservation system is integrated into the ETD service of California Digital Library.[27]

Beyond these external stakeholders, there are others who do not directly play a part in the implementation of ETD programs, although they have an impact on one or more aspects of ETD operations. For instance, ETD end users, both local and distant, provide input on how to search and use ETDs effectively and efficiently; and ETD funders (e.g., government agencies or private for-profit organizations) may greatly influence the embargo period of funded ETDs.

[19] MetaArchive was founded as part of the Library of Congress's National Digital Information Infrastructure and Preservation Program, http://www.metaarchive.org (last accessed 11-21-2012).

[20] MetaArchive has developed an organizational model and implemented a technical infrastructure based on LOCKSS software to preserve ETDs, http://muse.jhu.edu/journals/lib/summary/v057/57.3.skinner.html (last accessed 11-21-2012). Since 2008, MetaArchive Cooperative has partnered with NDLTD to undertake a preservation venture, an ETD dark archive designed specifically for ETDs in higher education institutions through the NDLTD/MetaArchive distributed digital preservation network.

[21] See Networked Digital Library of Theses and Dissertations, ETD Preservation, http://www.ndltd.org/resources/etd-preservation (last accessed 11-15-2012).

[22] LOCKSS (Lots of Copies Keep Stuff Safe) is an award winning, open-source digital preservation software released in 2004, http://www.lockss.org/about/how-it-works/ (last accessed 11-21-2012).

[23] See http://www.lockss.org/community/networks/ (last accessed 04-10-2013).

[24] DuraCloud is a cloud-based service developed and hosted by the nonprofit organization DuraSpace. It offers a simple and scalable cloud-based solution to preserve digital content in using multiple cloud service providers such as Amazon or Rackspace. It also allows users to replicate and access their digital content in the cloud. http://www.duracloud.org/faq (last accessed 11-15-2012).

[25] See Kimpton, Michele and Jonathan Markow, "Building community clouds to support access to scholarship," http://docs.duraspace.org/documents/DuraCloudEducauseFeb2012.pdf (last accessed 11-15-2012).

[26] See https://merritt.cdlib.org/ (last accessed 01-23-2013).

[27] See http://www.cdlib.org/cdlinfo/2011/12/06/uc-electronic-theses-and-dissertations-etds-now-have-preservation-and-access/ (last accessed 01-25-2013).

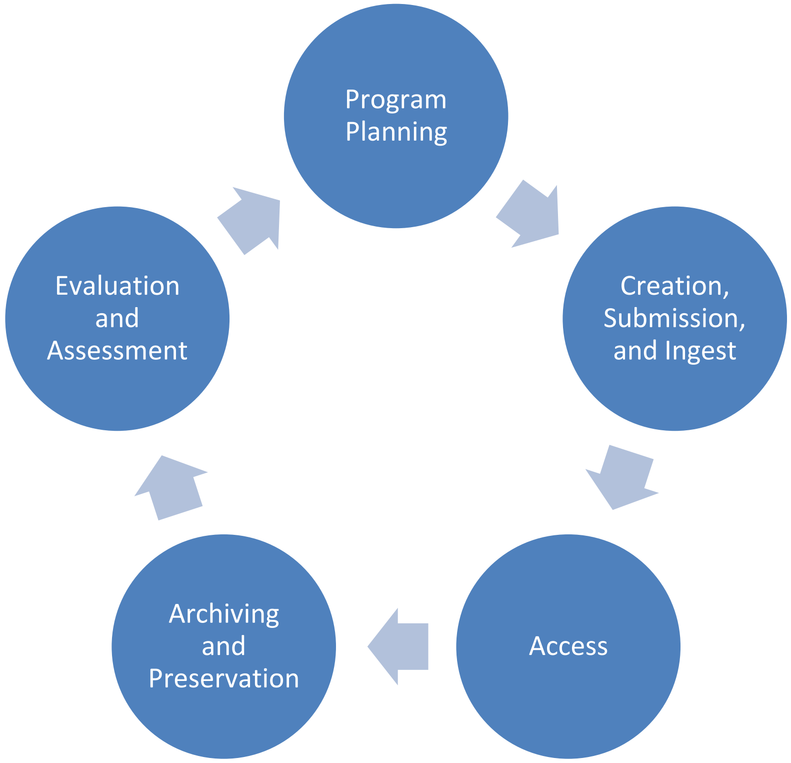

Program Planning
Creation, Submission, and Ingest
Access
Archiving and Preservation
Evaluation and Assessment

Figure 1-2. Stages of Implementing an ETD Program

The following sections attempt to outline the roles and responsibilities of the above stakeholders in ETD program management, although not all ETD programs receive participation from the potential stakeholders. The process of ETD program management involves planning, implementation, and assessment. The implementation process in particular covers several procedures such as ETD submission, ingestion, access and preservation. (see *Figure 1-2* for a diagram of the process)

1.3 Stakeholders' Roles and Responsibilities

This section details the roles and responsibilities of ETD planning committees and institutional administrators at the planning stage. At this stage, a number of internal stakeholders join the planning committee and undertake three primary tasks: providing a rationale for establishing an ETD program, advocating the program, and proposing an implementation plan.

1.3.1 ETD Program Planning

The first move towards instituting an ETD program is planning. Typically, a planning committee is formed to lead the work. The planning committee ideally consists of nominated or designated members from various internal stakeholders: the graduate school dean, the graduate school personnel, faculty members, graduate representatives, the chief information officer, the general counsel, the library dean, as well as the heads of the library digital initiatives, technical services, and reference units. The major responsibilities of the committee are charted below. Please see also *Guide to ETD Program Planning and Cost Estimation*.

1.3.1.1 Providing a Rationale for Establishing an ETD Program

The planning committee bears the responsibility for identifying the significance of ETD programs. Normally, the committee surveys the ETD professional literature, visits ETD websites, and consults peer institutions through face-to-face or teleconferencing meetings. By comparing the current local practice of handling paper-versioned theses and dissertations with ETD services, the committee may reach an agreement regarding the potential advantages of introducing an ETD program in general terms, as follows:

a. Increase the prestige of higher education institutions via open dissemination of high-quality intellectual output (Copeland and Penman 2004).
b. Provide greater visibility of underused graduate original research in the global academic arena.
c. Streamline and automate the processes from theses and dissertations' submission, distribution, to preservation.
d. Save resources that would otherwise be spent on printing, binding, shelving, storing, and circulating ETDs including through interlibrary loan services (Jewell et al. 2006).
e. Convey a richer message through the use of multimedia and hypermedia technologies, such as images, sound files, videos, datasets, and databases (Suleman et al. 2001).
f. Enhance graduate education (Fox et al. 1996).
g. Promote scholarly communication by sharing intellectual capital and supporting the open access movement.[28]
h. Promote developing digital libraries built upon collaboration among universities (Rodríguez 2006).

1.3.1.2 Advocating the Program

To gain support from every key sector of the academic community, the planning committee is responsible for promoting the value of ETD programs to the entire institution. Early involvement of the representatives from all concerned groups is imperative for the success of the project (Jewell et al. 2006). The planning committee is accountable for actively approaching stakeholders and engaging them in the establishment of an ETD program. Promotional and advocacy work includes not only the ETD program, but also the ideas of the open access movement, digital publishing, scholarly communication, and digital libraries (Jones and Andrew 2005).

The planning committee advocates the program to the university community and attempts to fully understand individual stakeholders' perceptions of the program. The committee should share the implementation details and seek comments and suggestions from the stakeholders. The committee is responsible for clarifying misconceptions, debating the pros and cons of the program, identifying possible areas of concern, and addressing issues raised by stakeholders. Advocacy problems often encountered include but are not limited to:

a. Lack of awareness of the importance of the program.

[28] See more on Open Access, http://www.arl.org/sparc/openaccess/ (last accessed 11-15-2012).

b. The lack of funds, trained staff, technical expertise, and infrastructure (Satyanarayana and Babu 2007).
c. Lack of university regulations and policies for ETD program implementation.
d. Copyright or intellectual property rights related issues (see also *Briefing on Copyright and Fair Use Issues in ETDs).*[29]
e. Perceived threats of plagiarism due to the free access of ETDs.[30]
f. Potential negative impact for future publication in journals and books.[31]
g. Concern over the quality of non-research degree theses (Bevan 2005).
h. General disinterest in or negative attitude towards changes.[32]

To communicate effectively with stakeholders, the planning committee needs to reach out actively to them. This may be accomplished through presentations about the ETD program at various campus meetings (e.g., administrative, departmental, and college faculty meetings); personal visits to faculty, staff, students, and administrators; the publication of articles in campus newsletters or newspapers; as well as invitations to ETD program operators from peer institutions to share their experiences and lessons with the local audience (Greig 2005).

1.3.1.3 Proposing an Implementation Plan

The planning committee takes on a further responsibility by drafting an ETD program proposal. In addition to specifying the program background, goals, and objectives and estimating associated costs and fees, the committee systematically investigates a range of core implementation issues:

a. What policies, regulations, and procedures with respect to ETD creation, submission, intellectual property rights, publication, and preservation should be made?
b. Where to host the ETD collection: in a third-party vendor's platform, a consortium of institutions, or a home-grown system?
c. What ETD submission, publishing, and preservation systems to adopt/develop: host platforms which allow for self-submission and publication such as DSpace, EPrints, Fedora, CONTENTdm, Ex Libris DigiTool, and VTLS Vital; or open source, in-house, or proprietary software ETD submissions systems that include graduate college review workflow, such as ETD-db, ProQuest ETD Administrator, Digital Commons, and Vireo.

[29] Copyright or intellectual property rights related issues are seen as a significant barrier often confronting institutions adopting ETD programs (Ghosh 2007). While students generally own the copyright of their work, home institutions or funding agencies may claim the rights, and commercial publishers own the rights when a copyright transfer agreement is signed.

[30] But some argue that Internet search engines, on the other hand, expedite detection of plagiarism (Yiotis 2008), which should mitigate the concern.

[31] Because ETDs that are freely available on the web may be considered prior publication. However, based on the earlier work (for example, An Investigation of ETDs as Prior Publications: Findings from the 2011 NDLTD Publishers' Survey http://vtechworks.lib.vt.edu/bitstream/handle/10919/11338/PriorPubs4ETDs2011Paper.pdf?sequence=3), "publishers do not necessarily see an ETD as an obstacle to a future publication" (Yiotis 2008).

[32] "Cultural issues are the most significant factors discouraging the adoption and development of ETD programs rather than technical issues" (Alsalmi 2008).

d. What suitable formats to accept for submission, access, as well as archiving and preservation (e.g., PDF and XML)? How to deal with ETDs with non-textual components? (see also *Managing the Lifecycle of ETDs: Curatorial Decisions and Practices*)
e. How to manage intellectual property rights, including fair use, copyright, plagiarism, access restrictions, and embargoes?
f. Which metadata standard to utilize in the cataloging of ETDs and whether to render metadata for harvesting?
g. What workflows to develop in regard to the life cycle management of ETDs?
h. Where to disseminate and access ETDs? Some options are library catalogs, institutional repositories, ProQuest, OCLC WorldCat, NDLTD, search engines, consortial systems, etc.
i. What access options to apply: worldwide open access, restricted access, fee-based access, or mixed access (Yale 2004)?
j. How to archive and preserve ETDs, including what media, formats, procedure, and strategies?
k. Whether and how to digitize retrospective theses and dissertations?
l. What IT infrastructure and technical support to employ?
m. What are the logistics with ProQuest or other external entities?

As a result of extensive and in-depth investigation, the planning committee proposes key decisions or plans, suggests a timeline and milestone events, and estimates the cost for program implementation. Most importantly, the committee stipulates the roles and responsibilities of key operational stakeholders at different implementation periods.

An ETD project must have full administrative support with an adequate budget (Rodríguez 2006). In the planning phase of an ETD program, institutional administrators have a significant influence on the adoption and development of an ETD program (Alsalmi 2008). The institutional administrators thoroughly review the proposal and make decisions to approve, reject, or modify the document. If the proposal is approved, the institutional administrators have the authority to (a) amend university-wide regulations, policies, and procedures for graduate degree completion and submission (Jewell et al. 2006); (b) allocate a budget and resources for the management and ongoing maintenance of the program; and (c) delegate responsibilities to individual university units/staff for ETD submission, access and preservation. Moreover, the institutional administrators are in a position to ensure that the key issues raised by internal and external stakeholders are addressed. Although institutional administrators are generally less involved during the public implementation phase, the institutional administrators monitor the implementation progress and may intervene in a particular step when deemed necessary.

1.3.2 ETD Program Implementation

ETD program implementation is a multipart procedure. To ensure the success of instituting this program, an ETD working group (usually including representatives of graduate schools and libraries) first conducts a pilot test. The pilot test limits the ETD operation to a particular academic period and/or unit(s). During the testing period the working group monitors the progress and evaluates the outcomes of the pilot (e.g., the faculty/student satisfaction level, the cost, and the efficiency of workflow). The most important job of this group is to identify the areas (e.g., ETD creation training and copyright support) that demand addition, adjustment, or elaboration. In addition, this group is responsible for verifying that the roles and responsibilities of the stakeholders are clear and that the stakeholders act in accordance with their duties.

With the experience from the pilot test and positive evaluation by the institutional administrators and planning committee, the ETD program enters a production period. The following section outlines the roles and responsibilities of key ETD stakeholders in four key stages: ETD creation, submission, and ingestion; ETD access; ETD archiving and preservation (with reference to the DCC Curation Lifecycle Model[33]); and ETD program evaluation and assessment.

1.3.2.1 ETD Creation, Submission and Ingestion

ETD creation, submission, and ingestion are a series of processes resulting in ETDs being electronically produced, submitted, cataloged, and rendered accessible through a digital repository. The chief stakeholders in this phase are graduate schools, graduate students, faculty, offices of general counsel, libraries, IT personnel, and several external stakeholders (e.g., ProQuest, NDLTD, and library consortia).

Graduate schools play a critical role in ETD creation and submission. They develop a body of ETD related policies and procedures, manage the ETD electronic submission process, and approve final manuscripts.

Developing ETD Submission Policies and Procedures

Graduate schools establish a set of relevant policies, articulating ETD guidelines, ETD templates, ETD formatting policies, ETD checklists, ETD embargo policies, etc. Graduate schools may adapt existing policies to incorporate ETDs, but two new policies need to be created for ETD submission format and embargo if it is permitted:

a. The submission format policy may restrict ETDs to a limited range of formats for both text and non-text files and develops a set of guidelines for formatting common content types. When making format decisions, a recommended practice is to consciously balance ease of production and access with ease of future migration/retention (Teper and Kraemer 2002).
b. The embargo policy withholds an ETD document or a portion of it for a defined period and/or limits it to certain types of users. When the embargo request expires, the ETD will become publicly accessible (see also *Guide to Access Levels and Embargoes of ETDs*).

Graduate schools can also team with external stakeholders (e.g., ProQuest), as well as various internal units, to establish a submission procedure, to create a submission form, and to determine a submission fee.

Offering Assistance for Students

Graduate schools may coordinate with IT personnel to set up a fully developed ETD website that distributes up-to-date ETD policies, format instructions, and other relevant documents. The ETD website can serve distance and on-campus students anytime and anywhere if an Internet connection is available. Graduate schools may be also responsible for creating tutorials and providing ETD consultation services

[33] See DCC Curation Lifecycle Model, http://www.dcc.ac.uk/resources/curation-lifecycle-model (last accessed 11-15-2012).

to graduate students as needed. Because ETDs particularly involve technical requirements and legal issues related to digital publication, it is a recommended practice that graduate schools conduct a series of workshops or training that describe ETD benefits, preparation, submission, access, preservation, student responsibilities, publishing checklist, institutional repositories, etc., at least one semester prior to graduation.

Administering the Submission Process and Approving ETDs

Prior to final submission, graduate schools verify the completion of submission and ensure that final ETDs are in conformity with all ETD requirements. They may be also accountable for administering the submission process and notifying other stakeholders (e.g., students, faculty, libraries, and ProQuest) in a timely manner regarding any decisions they have made. Lastly, graduate schools have the authority to either approve or reject embargo requests and final submission.

Graduate students assume the full responsibility of creating a research manuscript and converting the document into the required ETD formats. During the process, they may need a lot of support from their departments, the IT groups, the graduate school and other units on campus. The students may be responsible for submitting ETDs to designated repositories, depending on which submission method his/her home institution chooses (i.e., self-submission or mediated submission).

Constructing Theses and Dissertations

In terms of developing original research, students' responsibilities do not deviate greatly from writing a paper-formatted work. However, due to the nature of ETDs, students can add non-textual material (e.g., visual images, audio and video files, simulations, 3D visualizations, hyperlinks, and html) into their text-based documents assuming they respect other parties' copyrights. This is particularly useful for such disciplines as studio art, film and digital media, theatre, performing arts, and computer sciences.

Formatting Theses and Dissertations

(see also *Guidelines for Implementing ETD Programs – Roles and Responsibilities*)

Students are responsible for converting their manuscripts, including supplemental files, into acceptable digital formats (e.g., PDF, TIFF, MPEG, or AIF). In the conversion process, they must conform to any requirement imposed by their home institutions, such as embedding fonts, hyperlinks, and multimedia objects, removing security restrictions, and using non-proprietary file types. Should a student encounter formatting and conversion difficulties he/she needs to contact appropriate stakeholders for assistance.

Complying with Copyright Law and Making Embargo Requests

(see also *Briefing on Copyright and Fair Use Issues in ETDs*).

Many factors, such as the open access of ETDs, the use of a published journal article as a chapter of a dissertation, and the use of multimedia files, complicate the intellectual property rights surrounding ETDs. Therefore, students are responsible for understanding what rights they hold, what laws they need to abide by, and what liabilities and responsibilities they have when signing licenses in legal documents. For example, students should understand that a non-exclusive license does not relinquish the copyright of their work; instead, the license is to confirm that students retain the copyright of their research. Also, students hold the responsibility of appropriately using and citing copyrighted material (i.e., with permission, public domain, and fair use). Graduate school reviewers often require evidence of fair use analysis, including those in non-textual components, students should conduct a fair use analysis beforehand, seek permissions from copyright owners if fair use does not apply, and pursue help if they are uncertain. A recommended practice is to incorporate copyright information into ETD advice, services, and library instruction at the early stage of students' embarking on their research, as well as throughout their studies via various channels.

For concerns such as research containing sensitive data or works pending publication or patent, students may request an embargo (i.e., not publishing the work prior to a specific date) before final submission by following the local request procedure. Students who wish to extend the original period of embargo need to abide by local policies and submit a new petition before the manuscript reverts to open access. It should be aware that some institutions do not permit any embargo requests; therefore there is no embargo policy and processes in the ETD management.

Submitting ETDs

Unlike printed copies of theses and dissertations, ETDs reside in a virtual environment. Students are responsible for submitting their manuscripts to an appropriate web destination (e.g., an institutional repository or external publisher). At some institutions, in the online submission process, students are responsible for following instructions for a successful submission, including uploading approved final ETDs; paying a publishing fee for external stakeholders (if needed) and other mandatory fees associated with submission and graduation; choosing one of the access options; registering their copyright with the US government; and supplying metadata such as free-text keywords. If the documents are not compliant with the ETD submission requirements, students are requested to revise them for another submission.

Faculty members serve as graduate advisors and approvers of theses and dissertations. They are responsible for providing guidance on students' project development from initial proposal to defense (Indiana State University 2012). Faculty have primary authority approving theses and dissertations that will then be converted into required electronic format(s) and sent to graduate schools. Faculty members frequently share responsibility with the deans of graduate schools to approve embargo requests. In some institutions, faculty may participate in the review process of ETDs to certify that the final ETDs meet the expected standard of content and format (University of South Florida). In addition, through their own experience of digital publication and/or with concerns over their advisees' theses and dissertations, faculty may counsel students on intellectual property-related issues and suggest access levels of specific ETDs.

Offices of General Counsel are responsible for developing ETD related legal policies and offering legal counsel, in addition to providing legal principles for ETD programs.

Creating and Reviewing Legal Documents

Offices of General Counsel construct ETD-related legal policies, one of which is a non-exclusive distribution license that allows universities to openly deliver, reproduce, perform, and/or display ETD submission. The offices also undertake the task of reviewing and revising the partnership agreement between institutions and external parties. It is the responsibility of the General Counsel to advise the university community about intellectual property law and balance the needs of multiple stakeholders (Surratt 2005).

Providing Legal Services

The general counsel offices provide legal assistance for students, academic staff, faculty, and departments/units. For example, they may offer advice on using copyrighted material under the doctrine of fair use, obtaining explicit permission when fair use exemptions do not apply, publishing retrospective titles, and negotiating with journal or book publishers which rights to transfer. In addition, they may suggest tools and resources[34] for students to identify potential or unintentional copyright infringement. They may also supply students with templates for copyright requests to publishers and other copyright holders. Given the complexity of legal issues relating to ETD programs, the offices may be responsible for presenting legal workshops and providing individual consultation whenever needed.

Libraries are in a position to catalog ETDs, manage the processes of ingesting ETDs into libraries, handle retrospective ETDs, and prepare ETD preservation at creation. At some institutions, libraries are responsible for uploading final ETDs into local repositories.

Cataloging and Ingesting Born-Digital ETDs

Libraries have the responsibility of establishing an ETD automated cataloging workflow in local integrated library systems or institutional repositories. Cataloging and metadata librarians employ an ETD metadata standard (e.g., ETD-MS[35]) and develop a local ETD metadata set. The librarians provide technical, preservation, and descriptive information to create bibliographic records. To maximize ETD access points, cataloging and/or metadata librarians review author-supplied metadata and catalog these records according to library cataloging standards (e.g., LCSH[36] and presumably RDA[37] in the near future). Librarians, perhaps along with graduate school staff correct errors introduced by graduate student authors, make certain that special characters are represented properly, and most importantly, conduct name authority control and subject analyses (McCutcheon 2011).

In some cases, when ETD files with accompanying metadata are returned from outside agencies to institutions, libraries are responsible for the successful ingestion with the help of IT professionals who import and export ETD collections between systems and write scripts to transform metadata from one schema to another. Libraries supervise the transmitting procedure, harvest ETD metadata, map the

[34] See Fair Use Evaluator by American Library Association, http://librarycopyright.net/resources/fairuse/, also see Virginia Tech Fair Use Analysis Checklist, http://scholar.lib.vt.edu/theses/copyright/FairUseChecklistVT.pdf (last accessed 11-15-2012).

[35] See http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html (last accessed 01-22-2013).

[36] Library of Congress Subject Headings, see http://id.loc.gov/authorities/subjects.html (last accessed 04-04-2013).

[37] Allen B, Ashman. "A Brief Look at How RDA Is Being Used To Catalog Electronic Theses and Dissertations." *Kentucky Libraries* vol. 77, Issue 3 (2013): 16.

metadata to automatically populate bibliographic records in the MARC standard, conduct quality control on imported metadata, and ingest ETD files and associated metadata into local systems.

Digitizing and Ingesting Retrospectively Reformatted Theses and Dissertations

Digital initiatives departments are responsible for rationalizing a retrospective digitization project. This includes analyzing the necessity and expenses of the massive digitization of analog documents. The departments review legal rights with the general counsel or copyright expert before scanning without the permission of former students. Library administrators have the authority to choose whether to use local digitization services or outsource digitization to vendors such as the Internet Archive. If the library administrators decide to scan retrospective theses and dissertations in-house, digital initiatives are responsible for the digitization operation, such as providing recommended scanning equipment and software, defining digitization standards, OCR-ing text to enable full-text search, developing digitization and ingestion workflows, creating ETD metadata usually based on existing online bibliographic information, as well as ingesting digitized material into a repository. These departments also need to control the quality of final products (e.g., digitized items and cataloged metadata) because the scanning process may produce problematic results such as missing, duplicate or misplaced pages, data conversion, and file naming (Alsalmi 2008).

Libraries may play a much broader role in ETD creation, submission and ingestion, depending on the practices of individual institutions. Libraries may be solely responsible for, or participate in, ETD literature review, ETD creation support (e.g., the use of citation management tools), ETD training, and more.

IT personnel provide technical recommendations and support for ETD creation and submission. Campus IT experts suggest desirable formats to graduate schools. As diverse technology and software are typically employed during ETD creation, conversion, and submission (Hall et al. 2005), computer support specialists have the responsibility of preparing workstations and installing necessary applications, for example, MS Access, Cold Fusion, Java Scripting, LaTex, and Adobe Acrobat. The help desk staff may conduct workshops and create manuals to review technical details, such as how to embed non-textual material into PDF files.

The other significant responsibility of IT personnel is to prepare an electronic ETD submission system. After evaluating the suitability, functionality, interoperability, and sustainability of a possible software package (Copeland and Penman 2004), and considering in-house technical resources, expertise, the program's purpose, and funding, the IT personnel make a recommendation to implementation teams on whether to develop an in-house software or adopt an outside application (i.e., proprietary or open source). Chief information officers estimate the expenses that will be incurred (e.g., purchasing a new server). System administrators set up network infrastructure and a server where submission will take place. Afterwards, application specialists and/or program analysts develop a local system or install a third-party application for ETD submission along with some level of local customization. Once the submission system is ready, system administrators are responsible for setting up an LDAP or other authentication system, creating student logins, and developing programs to automate the submission process. Associated tasks include maintaining and administering the system.

ProQuest retains a noteworthy role in ETD submission and ingestion in the commercial sector. Cooperating with an institution doing business with ProQuest, ProQuest creates an ETD administrator website, a service that debuted in 2003.[38] The website serves as an electronic submission management site for students to submit ETDs and for graduate schools to review and oversee ETD submission.

To assist students with submission, ProQuest provides a range of services (e.g., a PDF conversion tool, the support of uploading multimedia files, and copyright registration) in addition to step-by-step instructions. In response to the open access movement, ProQuest offers an open access publishing option (with the specification of students' rights, access restrictions, expenses, etc.) that has been available since 2010, in addition to the traditional publishing model.[39]

ProQuest charges contributors service fees, including the publishing fee (if the contributor chooses open access publishing option), the cost of purchasing copies in a variety of formats, and copyright registration fee (if ProQuest, on behalf of the contributor registers the contributor's work with the US Copyright Office). Based on the agreement between an institution and itself, ProQuest is in charge of cataloging, archiving, and publishing approved ETDs at ProQuest. At the request of institutions, ProQuest sends ETDs with accompanying metadata and documents to corresponding institutions along with a submission report.

Networked Digital Library of Theses and Dissertations assumes an international leadership role in ETD initiatives. NDLTD developed an interoperable ETD metadata standard (i.e., ETD-MS) in 2001, based on the Dublin Core standard. ETD-MS sets up a guideline for cataloging ETDs because the standard metadata set is tailored to capture such information as committee members (advisors), degree names, and degree levels that are specific to ETDs. NDLTD also encodes the standard for cross walking with the MARC-21 standard and XML schema as well.[40]

Since adopting the use of OAI-PMH, NDLTD has been harvesting ETD metadata information on a periodic basis from individual NDLTD participating members into its international and central union ETD catalog.[41] Additionally, NDLTD collaborates with member institutions to create a submission process. It has also developed submission instructions for non-member individuals and/or occasional contributors (Networked Digital Library of Theses and Dissertations).[42]

Library consortia represent a joint venture to manage ETDs in a consortial setting**.** Library consortia not only serve as an ETD knowledge base and resource, but also deliver a range of services and undertake

[38] See http://www.proquest.com/en-US/products/dissertations/etd_administrator.shtml (last accessed 11-15-2012).

[39] See http://www.proquest.com/assets/downloads/products/open_access_faq.pdf (last accessed 11-15-2012).

[40] See ETD-MS: http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html#introduction (last accessed 11-15-2012).

[41] See NDLTD Union Catalog Project, http://www.ndltd.org/join/ndltd-union-catalog-project/ (last accessed 11-15-2012).

[42] See Networked Digital Library of Theses and Dissertations, "Submit ETDs," http://www.ndltd.org/submit (last accessed 11-15-2012).

initiatives of importance to consortial members. Generally, some of the services that the library consortia may provide during ETD creation, submission, and ingestion are as follows:

a. Develop a common platform to host ETDs for participants. For example, the OhioLink's ETD center is a central repository for ETDs from contributing universities and colleges in Ohio.
b. Create a submission and management system. For instance, TDL has created an ETD submission solution, Vireo, to handle the submission and management of ETDs with value-added features (e.g., tracking and managing the manuscript review process).[43]
c. Work with individual graduate schools to propose a publishing agreement and to develop a submission workflow.
d. Create a standardized ETD metadata set and provide crosswalks as well. For example, TDL created its own ETD MODS Schema (based on ETD-MS) and maps the schema to ETD-MS and TDL ETD DC, a qualified set of the Dublin Core standard.[44]
e. Support multimedia file submission.
f. Catalog ETDs. For instance, FCLA creates MARC records from METS format[45] metadata and puts them directly into the NOTIS integrated library system (Florida Center for Library Automation).[46]
g. Ingest ETDs to a central database for future indexing and publication alongside other digital resources.
h. Forward ETDs to ProQuest if needed. For instance, Florida Virtual Campus uses FTP to send ETDs to ProQuest (Florida Center for Library Automation).[47]
i. Provide a report of submission statistics as requested.

1.3.2.2 ETD Access

(see also *Guide to Access Levels and Embargoes of ETDs*)

ETD access is the process of making submitted ETDs visible, searchable and available in multiple venues. Graduate schools, offices of general counsel, libraries, IT personnel, ProQuest, NDLTD, library consortia, as well as access harvesters/facilitators share a joint effort in this regard. The job of these stakeholders is to optimize ETD display, discovery, and access, which in turn largely exemplifies various advantages of ETD programs (for example, access to the full text of ETDs where available).

Graduate schools and offices of general counsel primarily make policies relevant to ETD access and use, as well as manage access control.

[43] See more on Vireo, http://www.tdl.org/etds/ (last accessed 11-15-2012).

[44] See more on Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations, http://www.tdl.org/wp-content/uploads/2009/04/tdl-descriptive-metadata-guidelines-for-etd-v1.pdf (last accessed 11-15-2012).

[45] METS format is a standard XML schema commonly used by libraries.

[46] See Florida Virtual Campus, "FCLA support services for electronic theses and dissertations (ETDs)," http://fclaweb.fcla.edu/uploads/Priscilla%20Caplan/FCLA_Support_Services_for_ETDs.pdf (last accessed 11-15-2012).

[47] See previous footnote.

Generally speaking, graduate schools and offices of general counsel are accountable for developing ETD release/access policies, defining different access levels (ranging from immediate unrestricted access to closed access) and specifying how to apply for a particular access level. In practice, "most universities that encourage or require electronic ETD submission also encourage or require open access" (Suber 2008). The statistics gathered by Virginia Tech demonstrates the increased use of freely accessible ETDs compared to printed equivalents. However, it is reasonable to accommodate the need to postpone public access of some ETDs for a designated period.

An ETD end user license states how the end users of an ETD can use the ETD, for example, whether users possess the rights to distribute an ETD or derive works based upon the ETD. One practice is to license end users with one of the six Creative Commons Licenses (Perry and Callan 2006). Graduate schools and offices of general counsel share the responsibility of explaining ETD user licenses, as well as enforcing ETD access restrictions and other legal constraints, along with providing legal counsel for graduate students before and after graduation.

In conjunction with libraries, graduate schools specify the rights and responsibilities of stakeholders or personnel involved in access management. Meanwhile, they designate specific staff to authorize embargo requests, restrict or release ETDs for public access, monitor the embargo status of restricted ETDs, and work with students to monitor the final results of copyright requests, etc.

Libraries commonly perform a lead role in distributing ETDs, including retrospective copies through various channels. They also improve the visibility and accessibility of ETDs overall.

Distributing Born-Digital ETDs

For ETDs submitted directly to institutions, libraries are responsible for the timely availability of ETDs to the outside audience by accelerating the workflow from ETD submission to publication between graduate schools (or students) and libraries. For ETDs submitted to outside publishers (e.g., ProQuest or consortia), libraries handle the license that allows for campus access to external ETD databases. Sometimes when ETDs need to be ingested and published in local systems, libraries work with third parties to rapidly disseminate ETDs to local and remote users.

Distributing Retrospectively Reformatted Theses and Dissertations

Libraries have a duty to consult with legal officers about the appropriate access options for retrospective content, as there is a risk of copyright infringement. For example, former student authors may not allow the reproduction and open dissemination of their work, or unauthorized copyrighted material was used in the original theses and dissertations. Currently, one common practice is to publicly disseminate these digitized documents but to inform the student authors how to request access restrictions, because seeking permission for public access from previous graduate students would be prohibitively expensive (Perry and Callan 2006). To distribute retrospectively digitized theses and dissertations digitized in high resolution, libraries usually reduce the documents' sizes and then upload them to ETD publication systems (e.g., institutional repositories). Also, libraries may need to add an additional MARC field that contains the URL pointing to the web location of digitized material in the existing bibliographic records.

Assisting Users with ETD Search and Retrieval

Libraries undertake the task of developing instructions (e.g., ETD LibGuides) on where and how to effectively browse, search, and retrieve ETDs. To access internal and external ETD collections, libraries provide search and retrieval assistance as usual because ETD collections are actually part of the digital resources generated by institutions and managed by libraries. While interacting with faculty, students, and other ETD end users, reference departments may discover the limits of current ETD publication systems and suggest implementing value-added search functions (e.g., searching ETDs by committee chair). In addition, libraries share the responsibility with IT personnel to make certain that large ETDs and ETDs with multimedia components are accessible and reader-friendly to users.

Providing Multiple Access Points

To make ETDs discoverable both within and outside the university community, libraries are responsible for exploring an array of channels that may give the widest access possible to ETDs. In doing so, some recommended practices are:

a. Placing a direct link to the ETD portal on the front page of other ETD websites, institutional student portals, and educational portals (e.g., Blackboard).
b. Linking online bibliographic records directly to full-text ETDs.
c. Adding ETD collections to the list of library electronic resources.
d. Indexing ETDs with major search engines.
e. Registering institutional repositories containing ETDs with open access repositories (e.g., ROAR[48] and OpenDOAR[49]).
f. Exposing ETD metadata to aggregators who extract ETD information from OAI-compliant repositories.

One of the recommended practices is to become a member of NDLTD whose ETD metadata is then harvested into the NDLTD ETD union catalog as part of the global collection of ETDs.

IT personnel provide the technical infrastructure and support for ETD dissemination, search and access.

Ensuring the Visibility and Accessibility of ETDs

A primary task for IT personnel is to develop a user-friendly ETD publication interface where the ETD end user license is posted. To support textual and non-textual ETD content for proper viewing, the IT personnel are responsible for setting up software and hardware to appropriately present ETD contents for end users including those with physical disabilities or reading with handheld devices.

Another responsibility of IT personnel is to prepare the ETD collections for harvesting by enabling OAI or incorporating OAI in publishing systems. To protect students' intellectual property rights, these IT personnel may advise on employing a security mechanism such as encrypted digital signatures or watermarks when delivering ETD documents.

[48] See Registry of Open Access Repositories, http://roar.eprints.org/ (last accessed 11-15-2012).
[49] See Directory of Open Access Repositories, http://www.opendoar.org/ (last accessed 11-15-2012).

Enhancing ETD Searchability

In some institutions, IT personnel are in charge of tailoring ETD search and browse features to meet local requirements, typically based on the feedback from libraries and graduate schools. For example, program analysts may request a new criterion that conducts search by file format. At some institutions, to enhance ETD search, display, and retrieval, it is the responsibility of the offices to integrate third-party web discovery tools into institutional repositories or library cataloging systems.

ProQuest is the leading commercial publisher and distributor of theses and dissertations in the United States. It holds the most comprehensive repository of dissertations and theses in its ProQuest Dissertations and Theses database (PQDT). PQDT has grown from the former Dissertation Abstracts to PQDT Full Text that includes nearly three million searchable citations to dissertations and theses from around the world from 1763 to the present day, together with over one million full text dissertations that are available for download in PDF format (ProQuest).[50]

Access to ProQuest-based ETDs is generally by subscription only. Whether users can access the abstract, the citation, the first 24-page preview, and the full text of theses and dissertations where available depends on which particular ProQuest dissertations and theses service an institution subscribes to. ProQuest offers access and download to theses and dissertations including supplemental digital files for paid users, while providing online retrieval and copy ordering (print, PDF, or microform) services at the expense of non-authorized users. With the open access publishing model available since 2010, ProQuest now furnishes free access to ETDs at PQDT Open for any Internet users, provided that the students who submitted the ETDs opted to publish works for open access and paid with an additional charge of $95. Like degree-granting institutions, ProQuest manages access control and delays publishing some entries according to the embargo agreement between graduate student submitters and ProQuest.

ProQuest electronically delivers theses and dissertations through its information access and retrieval system. To improve the online search experience, ProQuest offers multiple options in searching, results display, and document view, including searching by language, looking up index terms, combining line search (which is designed to help build a precise search using operators to combine different fields that target users' search terms) (ProQuest),[51] sorting results by relevance or date, exporting/saving results, creating a formatted citation, and the like.

Networked Digital Library of Theses and Dissertations is dedicated to centralizing ETD resources and enhancing access to ETDs worldwide. NDLTD offers access to ETD scholarship contributed by participating institutions and consortia at no cost throughout the world. As of September 2012 with the support of individual institutions that have implemented the OAI protocol and registered the OAI interface, NDLTD has harvested more than 1.9 million records of ETD metadata into its seamless union from which users can access ETD data at the websites of individual institutions.

[50] See ProQuest, "ProQuest Dissertations & Theses Database," http://www.proquest.com/en-US/catalogs/databases/detail/pqdt.shtml (last accessed 11-15-2012).
[51] See ProQuest, "Overview," http://search.proquest.com/help/academic/webframe.html?Advanced_Search.html (last accessed 11-16-2012).

NDLTD has developed two tools (i.e., Scirus ETD Search and VTLS Visualizer) specifically for federated searching and browsing across multiple institutions simultaneously. NDLTD Scirus ETD search uses an older and more limited search interface, while VTLS Visualize provides a more dynamic and sophisticated discovery platform with such features as linking pages of results on a social media network and turning a query into an RSS feed. Besides these two search tools, NDLTD lists a variety of valuable search tools which focus on specific countries or regions, e.g., Australasian Digital Theses program and South Africa' National ETD portal.[52] All these search tools together deliver a wide range of ETD access points and thereby greatly promote the scholarly communication of ETD collections worldwide.

Library consortia generally offer access services for ETDs submitted through either individual or consortial submission systems. After indexing ETD contents, library consortia display ETDs at an exclusive ETD portal such as the OhioLink ETD center or distribute ETDs alongside other digital resources like TDL. In addition to providing federated search across ETDs from individual institutions, library consortia usually enable searching by institution. ETDs at member sites are normally freely visible and accessible via major Internet search engines. However, the accessibility of full-text ETDs is contingent upon each member's access policies. Consortia may be responsible for executing ETD access control, for example, temporarily or permanently removing ETDs with critical problems.

Access harvesters/facilitators use advanced search capabilities to provide access to ETDs. Access harvesters crawl the web and index ETDs from a broad range of open access ETD repositories, and then provide easy- to-search interfaces with improved features to expedite the process. These search engines notably expand the availability of ETDs in a channel beyond the traditional scholarly community.

Access facilitators working with libraries generally embed a discovery layer into integrated library systems (e.g., Millennium and Voyager) so as to provide seamless federated searching across the full breadth of library contents. They offer a range of sophisticated search functions, such as relevance ranking, faceted searching, social tagging, and reviews, which aid users in discovering ETDs quickly, easily, and effectively.

The OAIster harvesting group utilizes OAI-PMH to harvest and index full-text resources contained in open access collections worldwide, including over 450,000 full-text theses and dissertations. Since OCLC took over OAIster in 2009, the OAIster database has been integrated into OCLC via the WorldCat Digital Collection Gateway. ETDs contributed to the OAIster database are available from Worldcat.org and OCLC FirstSearch to base package subscribers as well as the OAIster website (OCLC).[53] OCLC WorldCat

[52] See more search tools at NDLTD, http://www.ndltd.org/find (last accessed 11-16-2012).

[53] See OCLC, "How to access the OAIster database," http://www.oclc.org/oaister/access/default.htm (last accessed 11-16-2012).

Dissertations and Theses provides access to more than eight million dissertations and theses. Many of these are available electronically, at no charge, directly from the publishing institutions.[54]

1.3.2.3 ETD Archiving and Preservation

ETD preservation ensures the long-term usability of ETDs regardless of changes in technology. It is critical that ETD stakeholders take into consideration digital preservation issues as they relate to lifecycle ETD management. ETD preservation is a complex and on-going process, involving such activities as data curation awareness, financial support, longevity of storage medium, preserving metadata, rights management, and technology obsolescence (Shearer, 2006). Institutional administrators, libraries, IT personnel, and some external stakeholders are the parties who manage ETDs for long-term readiness (see also *Managing the Lifecycle of ETDs: Curatorial Decisions and Practices*).

Institutional administrators have a critical role in the long-term commitment to ETD preservation. Due to the lack of general awareness towards digital preservation, institutional administrators are particularly responsible for clearly articulating the necessity of digital preservation policies for intellectual output, including ETDs, and incorporating digital preservation as part of the institutional strategic plan. This task is essential to preserve the primary student literature through garnering support from various stakeholders and securing stable funding even in times of economic difficulty. In addition, institutional administrators are responsible for adapting or creating the regulations and retention policies governing ETD preservation.

Libraries assume a leading and evolving role in terms of preserving ETDs in perpetuity. Because ETDs have become part of the library's digital collections, the task of archiving and preserving born-digital and digitized theses and dissertations historically has fallen to libraries. ETD preservation is a complex and difficult task as libraries deal with ever-changing technology, a growing variety of digital file formats, and the lack of well-established ETD preservation standards and best practices.

Advocating ETD Preservation and Developing a Formal Preservation Plan

One challenging task of libraries is to discuss an array of ETD long-term retention issues early in the program planning stage and throughout other implementation phases. Libraries, in particular, digital initiatives departments, are situated to form a university-wide digital preservation committee, to propose a long-term ETD retention plan, as well as to establish corresponding policies, workflows, and procedures. Libraries have a responsibility to examine the literature on best practices of digital preservation, to analyze the multitude of preservation choices (e.g., preserving ETDs in local systems, ProQuest's Vault, and/or ETD-specific preservation networks; open-source alternatives such as commercial solutions), and to recommend a comprehensive preservation strategy that goes well beyond simple ETD backup to full preservation.

[54] See OCLC, "OCLC WorldCat Dissertations and Theses (WorldCatDissertations)," http://www.oclc.org/support/documentation/firstsearch/databases/dbdetails/details/WorldCatDissertations.htm (last accessed 11-15-2012).

Organizing ETDs

Libraries should consider how to organize ETDs at the outset of implementing ETD programs with digital preservation in mind, because taking care of ETDs involves online storage, web delivery, and format changes. To avoid ETD collections becoming overwhelmingly cluttered over time, one recommended practice is to logically structure ETD repositories, standardize naming convention for files and directory structures, and control different versions of submissions and files created over time (Halbert and McMillan 2009).

Preserving ETDs in Reliable Media or Systems

With support from IT staff, libraries are responsible for storing ETDs in safe and reliable media or preservation systems, either online or offline, either onsite or offsite in multiple locations. Some example options are preserving ETDs in live servers, static storage media, and/or long-term preservation networks such as the distributed preservation network of the MetaArchive Cooperative. One associated task is to migrate ETDs from media to media over time.

Preserving ETD Contents

Libraries are responsible for preserving digital copies of scanned theses and dissertations. For the purpose of preservation, libraries usually archive the production files as well as the master files generated during the digitization processes. These files are typically large and uncompressed, which poses a challenge for storage space in parallel with increased budget demands. For born-digital ETDs, the evolution to ETDs that are solely or substantially composed of multimedia or other accompaniments may prove problematic for preservation and accessibility in future years (Jewell 2006). To retain the integrity of ETDs, libraries should make best efforts to preserve ETDs with critical components for complete readiness. For example, HTML files encapsulated within ETD documents must include all other referenced files (e.g., CSS and any other associated files) to properly execute the web-formatted contents.

Preserving ETD Formats

The level of format preservation support provided for an ETD is relevant to the file format in which it is created, as well as procedure-related decisions. Libraries bear the responsibility of preserving ETD formats, which includes forward migration, normalization and/or emulation (Caplan and Thomas 2006). Libraries should carefully weigh the advantages and disadvantages of possible preservation formats and consequently determine the most sustainable formats according to local requirements.

The ideal preservation formats are well documented, well tested, nonproprietary, widely distributed, and platform-independent (Fisher and Dollar 2000). In practice, as there is no single robust ETD file format for preservation, a number of institutions have decided to accept certain archival formats such as print, microform, PDF, and XML.

Considering the changing status of file formats and underlying support technology, libraries are responsible for converting or normalizing ETDs into accessible formats, as well as migrating them into succeeding formats upon obsolescence, in controlled, supported, or emulated systems for unimpeded access. The optimal format migration does not lose the original content, formatting, and functionality of ETDs (McMillan and Skinner 2009).

Preserving ETD Metadata and URLs

Often neglected areas of ETD preservation are ETD metadata and URLs. Libraries are responsible for extracting ETD metadata from ETD publication systems and saving it, including the descriptive, technical, and administrative metadata, on a periodic basis. Along with the development of PREMIS (Preservation Metadata: Implementation Strategies) whose charge is to define a set of semantic units that are implementation independent, practically oriented, and likely to be needed by most preservation repositories (Caplan and Guenther 2005), libraries should investigate the use of the PREMIS metadata schema and incorporate it as appropriate into digitization and ETD workflow processes. For complex digital objects, there is a growing need to use a metadata wrapper that contains all relevant ETD metadata in METS[55]/XML Schema and also provides pointers to individual elements of the objects. With respect to preserving ETD URLs, one of the favored practices is through a third-party service such as a Handle system[56] to assure the permanence and persistence of ETDs' web addresses.

Actively Preserving ETDs

ETD preservation demands active and continual actions for a full-scale ETD preservation service. Libraries are responsible for proactively implementing a preservation approach with dedicated staff and resources. Also, libraries should routinely assess the risks to ETDs' formats (e.g., deterioration or obsolescence), monitor the storage medium used, and check ETD fixity and completeness. Moreover, libraries need to record the actions taken on ETD preservation such as data replication, repairs, and reformatting in an ETD master registry file.

IT personnel play an active role to ensure that the software and underlying hardware enable better digital preservation treatment. They share the responsibility with libraries to examine preservation solutions in the market and to make a technical recommendation for the most appropriate strategy. Regardless if in-house or collaborative efforts between internal and external stakeholders are required, it is the responsibility of IT personnel to provide preservation infrastructure, sufficient storage space, and technical expertise. For instance, system administrators may migrate stored ETDs from one electronic storage/platform to another due to technical failure.

Furthermore, because "technology obsolescence seems to be the greatest technical threat in terms of digital preservation" (Salmi 2008), IT personnel are responsible for ETD reformatting. They investigate and supply ETD transformation hardware and software, as well as convert and migrate ETDs formats including associated formats (e.g., multimedia and hyperlink) into other readable formats. To safeguard ETD data integrity and avoid possible data loss, IT personnel are required to employ virus checking, fixity checks (e.g., checksum validation), versioning control, and other mechanisms as necessary.

[55] Metadata Encoding & Transmission Standards, see http://www.loc.gov/standards/mets/ (last accessed 01-25-2013).

[56] "The Handle System provides efficient, extensible, and secure resolution services for unique and persistent identifiers of digital objects." See http://www.handle.net/ (last access 01-25-2013).

Digital preservation services provide cost-effective and long-term preservation for a wide range of digital contents. Although the preservation practices of these services differ, stakeholders in this category have the following responsibilities in common:

a. Provide a technical preservation framework to archive and preserve digital collections including ETDs. For example, Amazon S3 provides a scalable backup and storage service to preserve digital information in its cloud platform for contributing parties on multiple devices across multiple facilities.
b. Assist institutions with content organization and ingestion into dedicated preservation systems. One example is the MetaArchive Cooperative which recommends organizing digital content in manageable and logical archival units, as well as providing a set of documents on how to ingest digital material into a distributed preservation network through developing plugins (xml files which tell web crawlers which file URLs to fetch and crawl) or through producing and submitting BagIt packages ("bags") for ingest.
c. Store ETD data in dark archival servers, keep content synchronized when preserved information is modified at the contributor end, and restore the files as needed.
d. Distribute redundant copies to multiple locations such as domestic and oversea networks. For example, LOCKSS technology enables replicating and storing data in multiple networked servers.
e. Transform formats when necessary for contributors. The DAITSS digital preservation repository software "implements active preservation strategies based on format-specific processing including, where necessary, normalization and forward migration."[57]
f. Provide online and real-time access to the preservation dark archives for a large variety of formats and content types at a designated web interface, depending on the agreement between the involved parties. For example, the streaming service of DuraCloud is designed to allow for easily embedding media at a contributor's web site directly from DuraCloud.[58]

1.3.2.4 ETD Program Evaluation and Assessment

(see also *Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs*)

ETD program evaluation is often overlooked for many reasons, such as no need, interest, or funding; no procedure in place; no responsible stakeholders or assessment teams; and no ETD-specific evaluation criteria, methods, instruments, and benchmarks. As a result, most institutions have not integrated program evaluation into ETD management and do not have an overall assessment of the ETD program, although a few institutions have limited evaluation activities (e.g., creating a web survey and counting the number of downloads). However, systematic evaluation plays a significant role for a newly instituted ETD program to receive continued support from various stakeholders and therefore to prosper over the passage of time.

ETD program evaluation is not a trivial process. The following section details the roles and responsibilities of internal stakeholders (i.e., institutional administrators, graduate schools, libraries, and

[57] Florida Virtual Campus, "Welcome to the DAITSS website!" http://daitss.fcla.edu/ (last accessed 11-20-2012).
[58] DuraCloud, "Services," http://duracloud.org/services (last accessed 11-20-2012).

IT personnel) in this management step. It briefly covers what to assess, what evaluation data to collect and analyze, what meaningful measures to employ, what instruments to create, and what importance a specific evaluation may produce (see also *Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs)*.

Institutional administrators relate the program outcomes to the proposed goals and objectives, based on the evaluation report(s) from individual internal stakeholders or assessment teams. Institutional administrators measure the overall benefits of the program (in learning, teaching and research) and gauge the return on investment at the macro level. For example, they evaluate whether the program raises the school research profile, promotes institutional scholarship, reduces the cost associated with processing and circulating paper-formatted work, and/or empowers students and universities.

Institutional administrators are also responsible for illuminating barriers among stakeholders and incorporating the evaluation results into decision-making processes. The decisions made for the further growth of ETD programs may include: aligning financial, human and technical resources; adjusting individual stakeholders' responsibilities; modifying institutional-level policies (for example, changing ETD submission guidelines from voluntary to mandatory).

Graduate schools, in particular, the administrative offices, are responsible for assessing submitted ETDs as well as submission procedures and support services.

Evaluating ETD Submission

For administrative purposes, graduate schools collect submission information from internal or external submission systems and various forms (e.g., graduate application forms, embargo request forms, and copyright owner request forms). Then they compile, analyze and report this information, such as the number of total submission, award-winning ETDs, embargoed ETDs, as well as submissions by discipline, format, graduation level, and graduation year. This assessment primarily evaluates the extensiveness of the ETD collection that can be used as trend data for the individual institution in comparisons to peer institutions (Digital Library Curriculum Project). It can also be used as a checkpoint for ETDs requiring follow-ups and for students needing further assistance.

Evaluating ETD Submission Procedures and Support Services

Graduate schools, together with other stakeholders (i.e., libraries, offices of general counsel, and IT personnel) are responsible for creating a suite of measurement instruments, for example, exit surveys, interview questionnaires, and evaluation forms. To gather responses from students, graduate schools may incorporate these instruments into one or two ETD submission processes. Three criteria (i.e., effectiveness, efficiency and satisfaction) can be utilized to assess these aspects of the submission operation.

Evaluation data graduate schools may collect include: the time, the steps, and the cost for students to accomplish the submission task in comparison with submitting paper copies; the creation and submission difficulties students have encountered; the satisfaction levels of faculty, students, and working staff towards the changed creation and submission practices of theses and dissertations; and the provision and timeliness of ETD support and training.

With the first-hand feedback information, graduate schools are responsible for modifying submission practices and reinforcing some areas so as to maximize the positive impact of ETD programs.

Libraries play a primary role for ETD program evaluation. Library administrators, digital initiatives, reference, and technical services departments are involved in evaluating the program from different perspectives. Libraries may be in charge of constituting an assessment group and preparing the final report. As ETDs are usually one of the digital resources provided to patrons and also one of the library digital initiatives, the standards, criteria, and methods for evaluating networked digital resources and services can be applied for ETD program assessment.

Library administrators evaluate program impact towards library services. Library administrators use the evaluation reports from individual departments within the library to assess whether the ETD program: advances digital library technologies; develops and populates a digital library that accommodates primary student scholarship; makes contributions to improve networked library services; and positions libraries in the trustworthy stewardship of institutional digital preservation. Moreover, library administrators assume the responsibility of reviewing the budget, timeline, output, and cost-effectiveness of digitization projects for retrospective theses and dissertations.

Technical services departments review and evaluate ETD cataloging practices. By cooperating with graduate schools, digital initiatives and outside stakeholders (e.g., library consortia and ProQuest), technical services departments have a duty to gauge the efficiency of overall ETD cataloging workflow and thereby create a well-established and organized procedure. These departments also assess how rapidly ETDs can be made available in publication systems, regardless of whether ETDs are hosted in institutional repositories or external systems. In addition, catalogers and metadata librarians conduct post-cataloging quality check to make certain that subject headings are designated, names are authorized, consistency are maintained, etc. User-supplied reviews or comments can be used to check how the created ETD metadata supports access and comprehension.

Digital initiatives departments generally evaluate ETD archiving and preservation practices. Digital Initiatives needs to periodically review digital preservation policies and procedures, ensuring that these are adequate and appropriate for implementing an ETD digital storage and preservation strategy. They should measure the reliability and effectiveness of the adopted approach, in conjunction with the assessment of the outcomes, such as the quality of digitized, converted, migrated formats and the loss (or damage) of data. Lastly, they evaluate the size of the ETD collection and estimate future storage needs.

Reference departments are responsible for conducting usability and accessibility studies of ETD collections and users. Because reference and subject liaisons have developed a close relationship with end users over time, reference departments are in the best position to conduct the assessment. They are responsible for developing and/or employing measurements such as observation of users, analyses of Google Analytics and ProQuest supplied data regarding ETD transaction, and the application of newer assessment measures (e.g., e-metrics and LibQual+) to collect, analyze, and report these qualitative and quantitative data. The areas reference departments may evaluate include user's information seeking behavior, as well as ETD usage patterns and accessibility.

Reference departments conduct user studies, including the awareness of ETD collections, different levels of experience and education, search strategies, motivation for searching, and satisfaction rate to understand users' ETD seeking behavior in completion of given search tasks. Reference departments may measure the usage of ETDs, such as the number of full-text downloads, unique visitors, page views, and search sessions (e.g., by domain, subject and access level). Interestingly, the ETD statistics analyses from Virginia Tech and National Digital Library in South Korea both reveal that ETD usage echoes the general academic calendar, which shows a strong academic orientation of ETD user groups (Zhang et al. 2001). In addition, these departments evaluate the accessibility of full-text and multimedia ETDs as well as service quality of ETD delivery.

The attractiveness of the ETD collections to the users and the ease of using technology all contribute to the overall usage (Fuhr et al. 2007). These user-oriented studies, combined with statistical reports, help capture a relatively accurate picture of user search behavior and usage, provide understanding of users' perceptions of ETD collections and websites, and identify gaps in ETD discovery and delivery, which in return may help librarians increase the sophistication level of end users of library resources and information services.

IT personnel take an active role with respect to ETD assessment. The technologists provide technical preparation beforehand for other units to measure various aspects of the program. For example, system administrators embed web-based surveys and Google Analytics for reference departments to gather and interpret user feedback.

In addition, IT personnel evaluate ETD programs from the technological perspective to ensure that ETD systems are fully operational, technically sustainable, and financially viable. The technical staff is responsible for designing and creating assessment methods, including developing surveys about technical assistance and functions, inviting IT experts to review technical components, and making use of system statistical reports (e.g., transaction log files and log analyzers). IT personnel conduct system-centered assessment and are typically concerned with system performance and usability. The areas IT personnel need to evaluate include:

a. Robustness and security of infrastructure and servers.
b. Adequacy and replacement of ETD long-range storage.
c. Placement of ETD disaster plans.
d. Usability of the core ETD systems (e.g., submission, publishing and retrieval, archiving and preservation, and ETD websites) where the interaction occurs among students, end users, ETD working staff, and administrators. This might include the assessment of the computer system, network performance, interface design (e.g., visual appearance and content organization), the handling of multimedia contents, and browsing and searching mechanisms. The navigation system is particularly necessary for an evaluation, because navigation disorientation is among the biggest frustration for web users (Jeng 2005).

e. Upgrade, improvement and migration of ETD systems, such as adding a student identity management layer into submission systems.
f. Scalability and interoperability of ETD systems.
g. Availability and currency of technical equipment and applications.
h. Service quality of technological support.

Such assessment information is important to generate better design recommendations, implement desired system features and optimize websites. IT personnel usually collect, analyze, and convey their findings to other stakeholders. (see *Figure 1-3* for a summary of key stakeholders by stage.)

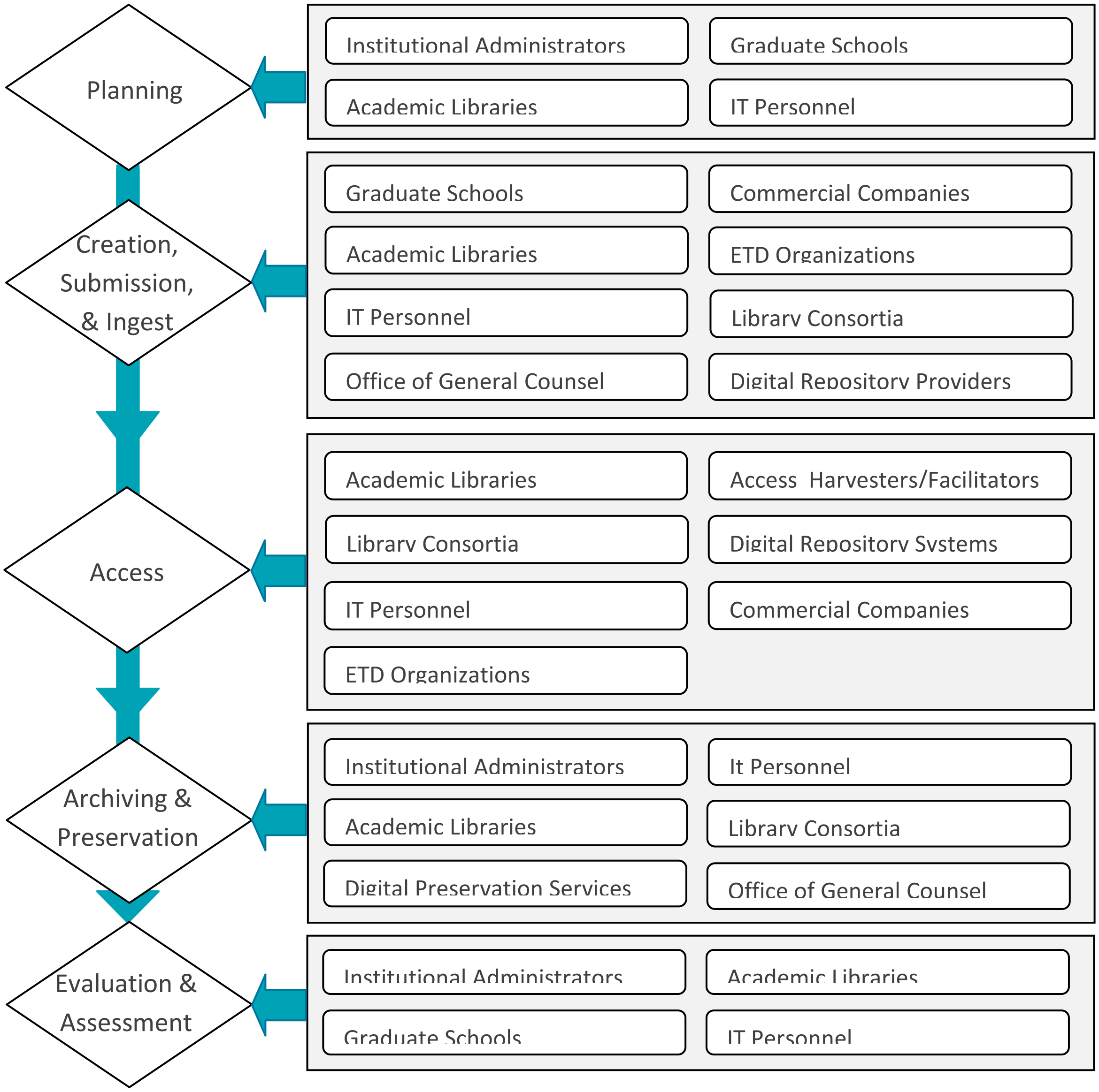


Figure 1-3. Key Stakeholders by Stage

1.4 Summary

Higher education institutions interested in pursuing an ETD program should understand the roles and responsibilities of the different types of stakeholders involved in the lifecycle management of the program. Issues these stakeholders may raise, such as ETD copyright, access restriction, long-term readiness, and support of a service gap, among other things should be taken into account and incorporated into the ongoing decision-making process. Having full-lifecycle management from the program planning stage, moving forward to the implementation and assessment stages, all while clearly specifying the roles and responsibilities of stakeholders and adopting best practices for ETD operation will help to ensure a successful program.

Bibliography

Alsalmi, Jamal. 2008. “Factors Influencing the Adoption and Development of Electronic Theses and Dissertations (ETD) Programs in University Libraries.” Paper presented in 11th International Symposium on Electronic Theses and Dissertations, Aberdeen, Scotland.

Beagrie, Neil and Maggie Jones. 2002. “Preservation Management of Digital Materials the Handbook." London: Digital Preservation Coalition. http://www.dpconline.org/component/docman/doc_download/299.

Bevan, Simon J. 2005. “Electronic Thesis Development at Cranfield University.” *Program: Electronic Library & Information Systems* 39 (2): 100-111. https://dspace.lib.cranfield.ac.uk/handle/1826/1353.

Caplan, Priscilla and Chuck Thomas. 2006. “DAITSS: Another Preservation Option for Electronic Theses & Dissertations.” Paper presented in 9th International Symposium on Electronic Theses and Dissertations, Quebec City, June 7-10. http://www6.bibl.ulaval.ca:8080/etd2006/pages/papers/SP10_Charles_Thomas.pdf.

Caplan, Priscilla and Rebecca Guenther. 2005. “Practical Preservation: The PREMIS Experience.” *Library Trends* 54 (1): 111-124.

Copeland, Susan, and Andrew Penman. 2004. “The Development and Promotion of Electronic Theses and Dissertations (ETDs) within the UK.” *New Review of Information Networking* 10 (1): 12. https://openair.rgu.ac.uk/handle/10059/46.

Digital Library Curriculum Project. “ETD Guide: Universities – The Assessment and Measurement Process.” Accessed November 20, 2012. http://curric.dlib.vt.edu/wiki/index.php?title=ETD_Guide:Universities_-_The_Assessment_and_Measurement_Process.

Ficher, Richard and Charles Dollar. 2000. “File Formats to Support Long-term Access to Electronic Records.” in *Managing Electronic Records Conference Proceedings* (Chicago: Cohasset Associates, Inc., 2000).

Fuhr, Norbert, Giannis Tsakonas, Trond Aalbert, Maristella Agosti, Preben Hansen, Sarantos Kapidakis, Claus-Peter Klas, Laszlo Kovacs, Monica Landoni, Andras Micsik, Christos Papatheodorou, Carol Peters, and Ingeborg Solvberg. 2007. “Evaluation of Digital Libraries.” *International Journal on Digital Libraries* 8: 21-38.

Fox, Edward A., John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer. 1996. National digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources. *D-Lib Magazine*. http://www.dlib.org/dlib/september96/theses/09fox.html.

Ghosh, Maitrayee. 2007. “ETDs in India: Towards a National Repository with Value Added E-theses Service.” Paper presented in the 10th International Symposium on Electronic Theses and Dissertations, Uppsala, Sweden. http://epc.ub.uu.se/etd2007/files/papers/paper-20.pdf.

Greig, Morag. 2005. “Implementing Electronic Theses at the University of Glasgow: Cultural Challenges.: *Library Collections, Acquisitions, & Technical Services* 29 (3): 326-335. http://eprints.gla.ac.uk/2295/.

Halbert, Martin and Gail McMillan. 2009. “Getting ETDs off the Calf-Path: Digital Preservation Readiness for Growing ETD Collections and Distributed Preservation Networks.” Paper presented in ETD 2009 12th International Symposium on Electronic Theses and Dissertations Pittsburgh, PA, June 10-13. http://conferences.library.pitt.edu/ocs/viewabstract.php?id=733&cf=7.

Hall, Susan L., Lona Hoover and Robert E. Wolverton J. 2005. "Administration of Electronic Theses/Dissertations Programs: A Survey of US Institutions." *Technical Services Quarterly* 22 (3): 1-17.

Indiana State University. 2007. “Responsibilities of Students & Dissertation/Thesis Chairs and Committees.” Accessed November 15, 2012. http://www.indstate.edu/gradexpress/td-responsibilities.pdf.

Jeng, Judy. 2005. “Usability Assessment of Academic Digital Libraries: Effectiveness, Efficiency, Satisfaction, and Learnability.” *International* Journal *of Libraries and Information Services* 55 (2/3): 96-121.

Jewell, Christine, William Oldfield and Sharon Reeves. 2006. “University of Waterloo Electronic Theses: Issues and Partnerships.” *Library Hi Tech* 24 (2): 183-196. http://www.emeraldinsight.com/journals.htm?issn=0737-8831&volume=24&issue=2&articleid=1558873&show=html.

Jones, Richard and Theo Andrew. 2005. “Open Access, Open Source and E-theses: The Development of the Edinburgh Research Archive.” *Program: Electronic Library and Information Systems* 39 (3): 198-212. http://hdl.handle.net/1842/811.

Kimpton, Michele and Jonathan Markow, “Building community clouds to support access to scholarship,” Accessed November 15, 2012. http://docs.duraspace.org/documents/DuraCloudEducauseFeb2012.pdf.

Lippincott, Joan K., and Clifford A. Lynch. 2010. ETDs and Graduate Education: Programs and Prospects. *Research Library Issues* 270: 6-15.

McCutcheon, Sevim. 2011. Basic, Fuller, Fullest: Treatment Options for Electronic Theses and Dissertations. *Library Collections, Acquisitions, and Technical Services* 35 (2/3): 64-68.

McMillan, Gail and Katherine Skinner. 2009. "NDLTD Preservation Strategy with the MetaArchive Cooperative." Accessed November 20, 2012. http://scholar.lib.vt.edu/theses/NDLTD/BoD200906/NDLTDPreservationPlan20090606.pdf.

Networked Digital Library of Theses and Dissertations. "Missions and Goals." Accessed April 10, 2013. http://www.ndltd.org/about/mission-and-goals.

Perry, M. and Callan, P. 2006. "Legal Protocols and Practices for Managing Copyright in Electronic Theses." QUT ePrints. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495722.

Rodríguez, Ketty. 2006. "Electronic Theses & Dissertations (ETD): A Literature Review." *Revista Puertorriqueña de Bibliotecología y Documentación* 8: 73-85. http://www.redalyc.org/redalyc/pdf/256/25600806.pdf.

Salmi, Jamal. 2008. "Factors Influencing the Adoption and Development of Electronic Theses and Dissertations (ETD) Programs, with Particular Reference to the Arab Gulf States." *Information Development* 24 (3): 226-236.

Satyanarayana, KV and B.R. Babu. 2007. "Trends in the Development of E-Theses in India: Issues, Constraints, and Solutions." Paper presented in 10th International Symposium on Electronic Theses and Dissertations, Uppsala, Sweden. http://epc.ub.uu.se/ETD2007/files/papers/paper-17.pdf.

Shearer, Kathleen. 2006. "The CARL Institutional Repositories Project: A collaborative Approach to Addressing the Challenges of IRs in Canada." *Library Hi Tech* 24 (2): 165-172. http://www.emeraldinsight.com/journals.htm?issn=0737-8831&volume=24&issue=2&articleid=1558871&show=html&PHPSESSID=6frv0c33uv5qo9uirrb5r3uot7.

Stanford University Libraries. "What is the LOCKSS Program?" Accessed November 15, 2012. http://library.stanford.edu/projects/lockss.

Suber, Peter. 2008. "Open Access to Electronic Theses and Dissertations." *Journal of Library and Information Technology* 28 (1): 25-34.

Suleman, Hussein, Anthony Atkins, Marcos A. Gonçalves, Robert K. France, Edward A. Fox, Vinod Chachra, Murray Crowder, and Jeff Young. 2001. "Networked Digital Library of Theses and Dissertations." *D-Lib Magazine* 7(9). http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html.

Surratt, Brian E. 2005. "ETD Release Policies in American ARL Institutions: A Preliminary Study." Texas A&M Digital Library. Accessed November 15, 2012. http://www.tdl.org/wp-content/uploads/2009/04/etd_release_policies.pdf.

Teper, Thomas and Beth Kraemer. 2002. “Long-term Retention of Electronic Theses and Dissertations.” *College & Research Libraries* 63 (1): 61-71.

University of South Florida. “Faculty ETD Checklist.” Accessed November 15, 2012. http://www.grad.usf.edu/inc/linked-files/ETD/FACULTY-CHECKLISTS.pdf.

Yale, Fineman. 2004. “Electronic Theses and Dissertations in Music.” *Notes* 60(4): 893-907.

Yiotis, Kristin. 2008. “Electronic Theses and Dissertation (ETD) Repositories: What Are They? Where Do They Come From? How Do They Work?” *OCLC Systems & Services* 24 (2): 101-115. http://www.emeraldinsight.com/journals.htm?issn=1065-075X&volume=24&issue=2&articleid=1728415&show=html.

Zhang, Yin, Kyiho Lee and Bum-Jong You. 2001. “Usage Patterns of an Electronic Theses and Dissertations System.” *Online Information Review* 25 (6): 370-378.

2 Guide to Access Levels and Embargoes of ETDs

Geneva Henry (George Washington University)

Topics Covered

- Reasons for students and institutions to restrict access to ETDs.
- Methods to restrict partial or complete access to ETDs.
- Benefits for students and institutions of not restricting access to ETDs.
- Methods of managing access restrictions from application to renewal.

2.1 Introduction

The transition from print to electronic theses and dissertations (ETDs) has led to increased scrutiny over who will be allowed to access the electronic versions and how widely they will be disseminated. With print-only, access to a thesis is necessarily delayed due to the time required to print, bind, and process the work for availability. The physical nature of the print also imposes restrictions on access since users are required to either purchase the work, go to the library of the institution where the thesis was published, or request it via interlibrary loan. These are all implicit barriers to broad dissemination of theses.

These impediments disappear when the works are submitted in their born-digital formats. The ability to widely disseminate the scholarship of an institution through the theses that are produced and to make the research available on the Web immediately upon submission of the final, approved thesis can prove advantageous to the newly-degreed student, the institution, and other researchers. In exceptional situations there may be concerns about making the research available immediately. For these circumstances, access restrictions may be imposed to address the concerns.

The purpose of this document is to examine access-related issues and provide guidance to ETD program stakeholders. These stakeholders include graduate student offices, graduate students, librarians, academic faculty, researchers, sponsors/funders, and other institution administrators who are responsible for making decisions about access to ETDs at their institutions or the institutions that were funded to do the research. This is a *living document* that will be updated as best practices continue to evolve around these issues.

This guidance document explores the issues related to access levels and embargoes of ETDs with the intent to help ETD stakeholders establish reasonable access policies for their institutions that will lead to consistent approaches and best practices for enabling access to ETDs. The document is structured to address the reasons for access restrictions, arguments against access restrictions, how restricted access policies compare across institutions, who makes the decisions about access, how restrictions are enforced, how stakeholders are informed about the release of a thesis from restrictions, implementing retrospective access restrictions, how theses are accessed, and a summary of the findings.

2.2 Reasons for Access Restrictions

Numerous issues can cause an institution to restrict access to an ETD. This section addresses the issues of how and when access happens, publishing concerns, the inclusion of sensitive data in an ETD, research sponsor restrictions, patent concerns, other types of concerns, and the policies for implementing access restrictions.

2.2.1 When and How Does Access Happen and What Does It Mean?

Access restrictions can be applied to an entire work or only parts of it. Embargoes are one form of access restriction whereby a thesis or parts of it are not available for a specified period of time; this is also referred to as a "publication delay" (United States Electronic Thesis and Dissertation Association 2012). In most situations, a thesis embargo is not intended to be permanent, but rather provide a means for delaying its public release, either in part or in whole (University of Kansas 2011). There may, however, be reasons for imposing a permanent embargo, though those are much less common. In addition to embargoes, redactions can be use to conceal specific information in a thesis even though the thesis is not embargoed. Redactions involve masking, or blocking out sections of the document that contain information that cannot be released. Access restrictions, more broadly, indicate that the full thesis or parts of it are not broadly available for some period of time, though they may be available to a limited audience such as members of a consortium or university (NDLTD 2010). There are a variety of access restrictions used by institutions to limit access to theses. The policies vary across institutions and the reasons for allowing access restrictions, as well as who makes those decisions, are many.

When a thesis or dissertation is available only in print, access requires more deliberate effort than when it is available digitally over the Internet. In discussing access restrictions for ETDs, it is useful to consider a comparison of how access functions in each medium (i.e., print vs. digital), and whether the ETD is a digitized version of a print document or a born-digital work.

Access to an ETD happens when the thesis is discovered online and either downloaded or opened for viewing. While it is natural to assume that an individual is doing this, it may also be a software program such as a "robot" crawling for theses or a software-harvesting engine collecting online works to aggregate them for dissemination through a common website. This did not happen when theses were available as print documents. In print, there were a very limited number of theses available to either borrow from a library or purchase from a thesis distributor such as UMI. (Publication of a thesis by a publishing organization, whether print or digital, has generally involved revisions to the original thesis to create a more refined work that would meet broader needs. This will be discussed in more detail later in this section.) Therefore, timing of availability, along with ease of distribution, are key issues that arise when considering how access will be provided to an ETD.

Many institutions have included digitized copies of their print theses in their ETD collections, making them more widely available than they were in print. However, retrospective digitization of theses and their inclusion in ETD collections must consider issues such as the institution's policy regarding the right to distribute these works in digital form, whether or not explicit permission is required from the authors to make them available online, what copyright license governs the work, and whether or not there is sensitive information included in the dissertation that should limit its distribution. Review of these and other issues identified by the institution may lead to a decision to restrict the availability of a retrospectively digitized thesis until the concerns are properly addressed.

With born-digital ETDs, many of the concerns that may be present with retrospectively digitized theses can be addressed from the outset of implementing an ETD program; these are discussed in detail in later parts of this document. Once online, the metadata that describes a thesis can be picked up by Internet search engines and readily discovered by anyone on the Web. Thus, timing of the dissemination of the research is significantly faster, offering the researcher almost immediate visibility and highlighting the type of research that prospective students might expect to engage with at the institution that issued the degree.

The role of traditional library catalogs in discovering an institution's theses is also worth considering. With print theses, the institution's library catalog included a record with the full bibliographic information for each thesis that was produced. These records have been and continue to be shared with both national as well as global catalogs such as WorldCat (OCLC 2012). With ETDs, the practice of including a record for each ETD, or a link to its digital version if it was digitized from print varies across institutions. Some institutions will include only a single record for the entire ETD collection, identifying the online location where all ETDs for the institution can be found while others are much more diligent about continuing to create item-level records for every thesis, regardless of its format, with links to the online versions. Without an item-level record, researchers using the catalog search interface for discovery of relevant resources will have more difficulty discovering them. For print theses that have no online representation, their discoverability is more limited for researchers who rely on Internet search engines such as Google Scholar to discover scholarly resources worldwide. While a MARC catalog record may exist and be available through a site such as OCLC's WorldCat.org, searches that use vocabulary that is not recorded in the catalog record will likely result in some relevant theses not being identified.

Catalog records for ETDs and their metadata record counterparts are not always consistent. Depending on the standard being used for each and the local practices for including some specific information such as department name and advisors, as well as the subject or key word terms associated with the thesis, the discoverability may vary for a user even though the catalog record is available online. Discovering either the metadata record or the catalog record for a thesis helps a user, but if the catalog reference does not provide a link for the digital version, the user may be disadvantaged in being able to access it.

In addition to the catalog and metadata records for an ETD, the thesis may reside in multiple repositories. Subscription databases such as the ProQuest Dissertations & Theses database (ProQuest LLC 2012) limit access to most of the theses they hold to subscribers only, unless the author has paid for their thesis to be made available open access. The Networked Digital Library of Theses and

Dissertations[1] enables discovery of the theses of participating institutions through a union catalog that directs access to the hosting repository. Many institutions now provide online access to their ETDs via their institutional repository or a repository dedicated to managing ETDs.[2] Students can also elect to have their thesis that they submit to a company such as ProQuest distributed by third parties, resulting in further distribution by sellers such as Amazon.[3] Many institutions still require their students to submit their theses to ProQuest in the belief that it is the authoritative database of all theses and dissertations in the US. Graduate students, however, are beginning to take issue with such requirements, sometimes arguing quite articulately that there is not a benefit to them to have their thesis in the ProQuest database of theses and dissertations (Clement 2012). There is not currently a single repository that serves as the official repository of all theses, though several institutions have expressed a strong desire to have one identified ("LISTSERV 16.0 – ETD-L Archives – October 2012" 2012).

Discovering that a thesis exists through metadata or a catalog record is different from having access to the actual work. Catalog records may fail to provide a link to the electronic thesis. Even though a descriptive record may be available to show that the thesis does exist, any access restrictions or impedances on an ETD will limit its distribution. Embargoes for a specified period of time, limited distribution to campus IP addresses, requiring a subscription to a database and other restrictions are ways of limiting access to theses. Policies for allowing theses with access restrictions to be discovered via a metadata record vary across institutions, with some institutions making their metadata and catalog records for the works available while others hide the record until the thesis can be distributed. While many may view a metadata record as independent of the thesis and argue for its availability even when a thesis is embargoed, there may be very valid reasons for not making the record available. If verification of a student's degree completion and successful completion of a thesis is required, the administrative units of the institution should be contacted directly rather than relying on the discovery of a metadata record for the student's thesis. For individuals searching for a thesis they know has been written and successfully defended, it can be very frustrating to not be able to find any references to the work.

2.2.2 Publishing Concerns

The definition of whether or not an ETD is considered "published" once it has been submitted in fulfillment of the degree requirements remains unclear. This has led to questions and practices around ETD embargoes specifically related to publishing. The scholarly publication of work based on a thesis almost always takes a different form than the thesis that was submitted in fulfillment of degree requirements (Ramirez et al. 2012). In the past, publishers were more likely to express reluctance to enter into a publication agreement with an author whose thesis was available online, though policies

[1] See http://www.ndltd.org/.

[2] Examples of institutions with dedicated ETD repositories include Virginia Tech (http://scholar.lib.vt.edu/theses/) and the University of Western Ontario (http://ir.lib.uwo.ca/etd/).

[3] For example, see http://www.amazon.com/s/ref=sr_st?bbn=173514&keywords=dissertation&qid=1354292442&rh=n%3A283155%2Ck%3Adissertation%2Cn%3A!1000%2Cn%3A173507%2Cn%3A173514%2Cn%3A227191&sort=daterank.

around this were often undocumented by various publishers (Seamans 2003). More recent surveys of publishers are finding that they are now more likely to view an ETD as a pre-print rather than a previous publication since the editorial work needed to publish it in a form that is considered a finished scholarly work is often significant (McCutcheon 2010; McMillan et al. 2011; Ramirez et al. 2012). For works such as creative writing or chemistry publications, however, publishers are likely to be more reluctant to publish the work if the thesis on which it is based is already available online. If planning to publish their work with a particular publisher, students should contact the publisher and confirm that an openly accessible copy of their thesis will not preclude the later publication by the publisher. Several institutions with creative writing programs still will not allow those theses to even be submitted electronically, insisting on print as a way to minimize the distribution of the thesis prior to its publication through a publishing house.

Since publisher policies and practices are still perceived as ambiguous, some schools are choosing to restrict access to recent theses to allow students time to publish their work with a recognized press that will further their publication credentials. Many, however, realize that making their theses available as open access documents increases their visibility and can lead to an increased number of citations to their work.[4] Publishing editors, however, may question the validity of citations to a thesis as a valid work if they consider the thesis to not be published. There is some evidence to indicate that editors will reject a citation to a thesis if it is not available online, referring to it as an “unpublished work” (Olson 2012). Other editors have stated their belief that ETDs will generally not be cited, only publications that have been through a publisher’s peer review process will be recognized and cited as reputable sources (Ramirez et al. 2012). As the transition to ETDs from print theses continues, many of these citation issues are likely to be resolved as both publishers and institutions gain a better understanding of ETDs. One measure that can help encourage ETD citations is to include a recommended citation format to the work in the thesis; this may be helpful to users who may not be familiar with how to cite an ETD so they can include the citation in their own works.

While publishers are increasingly regarding ETDs as separate from a final, peer-reviewed publication, issues remain about the impact of openly accessible ETDs on publications. One concern that publishing editors have is the ability to conduct a fair, blind review of a work to be published that is derived from an ETD. It is very easy to search online for a thesis, especially if the work to be published retains the title used for the thesis, thereby allowing reviewers to know who wrote the work they are peer reviewing. This can introduce bias into the refereeing process (Ramirez et al. 2012).

Another concern with open access ETDs relates to library acquisition practices in working with approval plans. The “dissertation factor” that some academic libraries apply to their profiles for acquiring new books may exclude theses that are available as ETDs, with acquisition librarians hesitant to purchase the peer-reviewed publication if it will be almost identical to the original ETD that is available through the

[4] See the guidelines document on *Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs* for a more detailed discussion of increased usage of open access ETDs.

institution's ETD repository. Publishers note that this behavior can lead to their reluctance to publish an already-available ETD if it means that no one will buy the book (Ramirez et al. 2012).

2.2.3 Sensitivity of Data and Sponsor Restrictions

Research involving sensitive data such as classified government information, industry trade secrets, or personal information that could compromise individual identities may be a reason to restrict access to a thesis. Many times, the company or agency that sponsors this type of research requires that any publication, including theses, resulting from the research be restricted in whole or in part. Medical research involving patient data must comply with federal policies such as the Health Insurance Portability and Accountability Act (HIPAA) to ensure that patient privacy is safeguarded. Research sponsored by federal funding agencies will often receive more oversight at an institution to safeguard the rights of individuals involved in experiments, ensuring that the researcher has received approval from the Institutional Review Board (IRB) before any research is conducted. The student and their advisors should be aware of any sensitive information included in the thesis and what limitations, if any, there are on publishing the research results for broad dissemination. Sensitive data can often be redacted to block the information from being released without embargoing the full thesis. Researchers working in subject areas that are likely to rely on sensitive data should clearly think through the impacts on the thesis and how best to handle it so that the student's work can enjoy the greatest possible visibility. Virginia Polytechnic Institute provides an exemplary policy for graduate students who may have sensitive data or sponsor-related restrictions associated with their theses, advising students to seek a "pre-research review of their thesis or dissertation plans with the sponsor whenever there is a possibility that certain findings might be subject to embargo," (Graduate School, Virginia Tech 2013).

Sponsors of research may require the published results to be made available as open access (e.g. the National Institutes of Health and the Wellcome Trust) documents, though this may not apply to theses since they are not peer reviewed journal articles (National Institutes of Health (NIH) 2009). Other sponsors may require that certain information not be disclosed for some period of time to protect any discoveries that may be either strategically sensitive in nature (e.g. military concerns) or beneficial to the sponsor in some way (e.g. chemical discoveries that can be commercialized). The terms of restriction should be clearly identified by the sponsor at the time the funding is provided for the research.

If possible, redactions of specific information in a thesis can be used to make the overall thesis less objectionable for release, avoiding the need for a full embargo. This, however, may require the institution to manage two versions of the thesis: the redacted version for public viewing and the full, original version that does not have redactions. Both will need to go through the digital curation process and the institution will be responsible for managing access accordingly. When theses are provided to third party providers such as UMI/ProQuest, additional complications such as required updates to microform images for redaction will require students to bear an additional financial burden to have their thesis redacted in all versions that the third party vendor maintains.

Graduate students may be supported by multiple grants in doing their research. Any restrictions by the sponsors must be clearly identified and reconciled to ensure that there are no conflicts in policies regarding availability of the thesis. If the sponsor has stipulated that they must review the thesis prior to its publication, embargoing the thesis for six months can often provide sufficient time for this required review (Duke University Graduate School 2013). Graduate schools that define policies, checklists and submission workflows can help to ensure that any sponsor restrictions are met by including information that allows the student to identify the sponsors and any restrictions at the time the thesis is submitted.

2.2.4 Patents and ETDs

Research discoveries may be eligible for patent claims. Since the patent application process can take some time, it is not unusual to request an embargo on a thesis until the patent request has been filed and the claim published. Many institutions are supportive of accepting theses in fulfillment of graduation requirements even though a request has been made to embargo its availability while awaiting a pending patent application. When a patent application is filed, an application number is assigned to the request. If this is done online, this number is available right away; otherwise, a paper filing will receive an application number within eight weeks of submission. The average time for the US Patent and Trademark Office to process the application is approximately 24 months ("Questions and Answers – USPTO - USPTO" 2003). If filing for a patent outside of the United States, the researcher must be cognizant of other countries' laws regarding patents since each nation has its own policies.

Until very recently, patent applicants have had a one-year window to apply for a patent after the idea or invention has appeared in a publication. This online availability of a thesis could protect a patent applicant in the event that someone else filed for a patent on the same thing, but after the thesis was made available online. An ETD that is available online may be considered a prior publication, thus the author will have one year from the time it is made available to apply for a patent for ideas or inventions discussed in the thesis. To allow an author more time to apply for a patent, institutions may choose to provide an embargo that will lengthen the time available to the author. The embargo acts to delay the start of the one-year window since it does not begin until the thesis is published (Duke University Graduate School 2013). Changes in US patent law effective March 2013 recognize *first to file*, but the one year of protection following disclosure of the idea will still protect US patent applicants (USPTO 2013). The move to first to file aligns with the copyright policies of most other countries in the world. The one-year grace period that US patent filers have, however, is not common in other countries in the world.

2.2.5 Other Reasons for Access Restrictions

There are numerous additional reasons a thesis author or institution may have to restrict access. Ethical concerns such as not violating cultural norms or beliefs may factor into access restriction decisions. For example, theses that examine writings by ethnic groups whose cultural beliefs may have gender-specified audiences for particular works may need to be restricted out of respect for those values. Known international prejudices that may be sparked by a thesis subject and that could lead to violence or harmful transgressions against a group or individuals may have their access restricted to prevent such actions. As the political landscape continues to change, issues that were once regarded as non-controversial may become hotbeds of tension and vice-versa. This shifting landscape may cause institutions to reconsider embargoes of theses where the subject becomes polemic. Theses that involve prurient materials may be embargoed if images that are included, though presented in a scholarly context, are deemed objectionable to some. A religious school may have stricter policies in this regard than secular institutions. Occasionally, there may be individuals who have, for legitimate reasons, requested and received permission from the institution to restrict all information about the individual. In these situations, an institution presents no externally available information to indicate that the individual has any affiliation with the college or university; this may include any information about a thesis.

Students, their advisors and the institution's general counsel will be in the best position to assess risks and determine whether or not the information can be redacted, the thesis embargoed, or if it can be made fully available. Identifying theses that should be restricted after they have been made available online will likely happen after it has been discovered by someone who contacts the institution and provides sufficient information for requesting that the information be removed from online availability. The final decision should rest with the institution and their policies regarding sensitive information and is likely to be assessed on a case-by-case basis. The author of the thesis should be informed of this changed status and offered an opportunity to provide a response to the concerns that have been raised. The above are examples of legitimate reasons for granting an embargo for a thesis, but it should be stressed that these instances are the exception rather than the rule.

Regulations regarding export controls can also lead to access restrictions on theses. This is more common in science and engineering subjects. United States export control laws are designed to protect the national security and foreign policy objectives of the US (Department Of State. The Office of Electronic Information 2011). As with ethical concerns, the changing political environment can result in changes to these regulations, with new controls enacted on content that was previously clear and old concerns dropped, enabling their open dissemination. While students may not readily be aware of the materials that would be subject to export control regulations, there are generally offices at research universities that confirm compliance with export controls on any grants or contracts.

Some institutions may restrict access to theses to either on-campus use or via login. This practice was more prevalent prior to 2012; many institutions that were uniformly restricting ETD access have now made their theses available as fully open access documents. Access may continue to be limited at institutions that are starting ETD programs and have concerns about immediate access to theses. Restricting the access to campus-only use mimics the print availability when the theses were maintained on library shelves. Institutions may also charge for outside access as a means to recover costs associated with the ETD program. MIT, for example, allows access to a non-printable version of their theses online, but restricts access to the printable download version to MIT users or those who are willing to purchase the PDF (MIT Libraries 2012). The guidance document, *Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs*, provides helpful information that can guide the decisions about whether or not there is a benefit in restricting access to an institution's ETDs.

2.2.6 Consistency of Institutional Access Restriction Policies

Determining access restrictions can and does happen at many levels. Institutions that are part of a multi-campus system may choose to have a consistent policy for how access restrictions are handled at all campuses or may allow each individual campus to set their own policies for restrictions. Within a campus, the policies on how access restrictions are made can vary widely since not all campuses have a centralized process for handling theses. At institutions where each individual department or school handles the thesis submission process, there may be a wide variance in policies for restricting access. As noted above, there are many reasons for determining whether or not there should be any restrictions placed on a thesis and the level of restriction to apply.

A survey of higher education universities in the U.K. found that those institutions vary widely in how embargoes are applied. The majority of access restrictions involve full embargoes of theses rather than redacted versions. Some institutions automatically embargo all theses while others permit embargoes for up to 10 years, with very few permitting embargoes beyond that. The most common practice is to allow embargoes for short, but renewable, periods of time, e.g. 1, 2, and 3 years. While reasons for embargoes varied, the most common reason stated for restricting access to a thesis was the presence of sensitive information (Education et al. 2012).

Online archives of scholarship have commonly been referred to as “dark,” “dim,” or “light,” connoting the degree of openly available access to the archive’s content (Kenney et al. 2006). This terminology has been applied to ETD repositories, as well to indicate whether or not the theses are openly available for free access worldwide. Dark archives do not provide any external access to the content, but do store the theses for long-term preservation. Light archives, by contrast, make the theses available to everyone. Dim archives provide restricted access to theses managed in the repository and the conditions for access can vary significantly. As discussed above, there are many approaches and reasons for limiting access to theses that would lead an ETD repository to be classified in the “dim” archive category.

2.2.7 Where to Provide Information about Access Restriction Policies

Institutions generally provide guidance about access restrictions to graduate students and their advisors through policies published on websites that provide information about thesis requirements. In very large research institutions, this may be available at each individual school or college rather than for the institution at large. Smaller institutions, however, may have their processes more centralized, providing institution-wide policies about theses and dissertations. The Web readily supports cross-referencing to appropriate websites, facilitating finding information about the appropriate policies and procedures for a graduate student. Departments, schools, colleges, and overall institutions that have policies regarding graduate education and thesis requirements should link to appropriate sites within their own institution to help students navigate the procedures they need to be aware of, including access restriction policies, in successfully submitting their ETDs.

In addition to the academic units and graduate student offices, the offices of research, compliance and the library are other units that should provide information about access restriction policies and how ETDs will be managed. Since these policies are changing as ETDs become more prevalent, it is important that graduate students check the sites frequently for updates and that the institutions keep the students informed of any changes. Sites that support automated updates, such as RSS feeds or change alert messages, would be one way to proactively alert the students to updates in the policies.

Making sure that both the students and their advisors are aware of the policies is part of the education process that should occur prior to the start of the research. Among the key issues, stakeholders need to be knowledgeable about the following topics when making access restriction decisions:

- Costs and benefits associated with access restrictions.
- Which restrictions, if any, are possible.
- Processes that must be followed in requesting a restriction.
- Responsibilities associated with embargo renewals.

An early education process covering these and other relevant topics for making decisions about the appropriate level of access for a thesis will result in clearer and more consistent decisions. While the education about these issues is intended to help students make an informed decision, there is likely to be as much impact on the faculty and other administrators who work with the students in providing overall guidance throughout their thesis process.

2.3 Arguments Against Access Restrictions

This document has focused on various types of access restrictions and reasons for restricting access to ETDs. There are, however, great benefits to not restricting access to ETDs. The benefits of not restricting access are also discussed in the *Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs* guidance document. In this section, a few reasons are highlighted for supporting open access to ETDs.

Increasingly, institutions and funding agencies are requiring open access to publications produced by their researchers and faculty. The Registry of Open Access Repositories Mandatory Archiving Policies (ROAR) tracks the number of open access mandates that have been passed, presented here as *Table 2-1* (University of Southampton 2012). *Figure 2-1* from the ROAR site illustrates how open access policies have grown over the past ten years, though the graphic does not capture the change in open access mandates for theses.

Open access thesis mandates started to appear in 2008, with one mandate, followed by 41 thesis mandates worldwide in 2009. In 2010, 35 more were added, in 2011, 14 more, and in 2012, 7 more thesis open access mandates were registered (University of Southampton 2012). This increasing trend towards open access mandates suggests that institutions should be mindful of the shifting culture that expects scholarship to be openly available. Restricting access to the scholarship produced by student theses runs counter to this trend.

Mandate Type	Number
Institutional Mandates	161
Sub-Institutional Mandates (e.g. School or College within an institution)	34
Multi-Institutional Mandates (e.g. multiple campuses in a university system)	4
Funder Mandates	54
Thesis Mandates	98
Pending Mandates	24

Table 2-1. Open access mandates as of January 2013

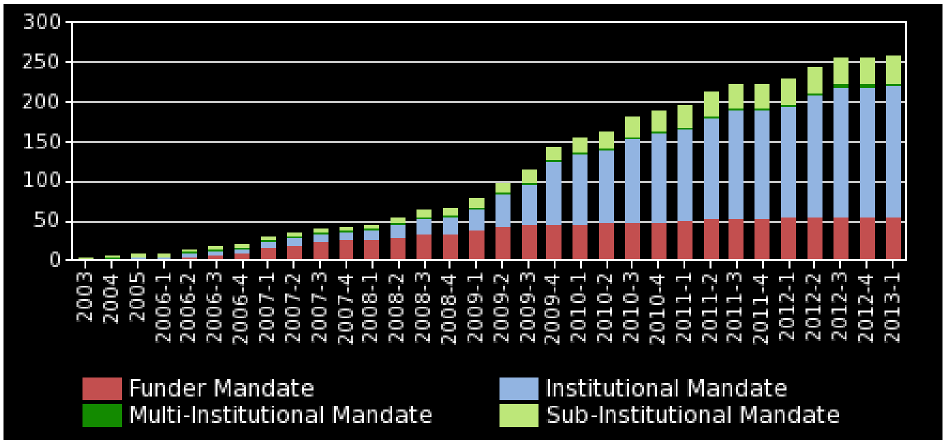


Figure 2-1. Growth of open access mandates 2003 - 2013

Concerns about the ability to publish a thesis have been shown to be unfounded, as discussed in section *2.2.2*. In surveys with journal editors and university press directors, ETDs are regarded more as pre-prints, requiring significant revision prior to acceptance as a publication (Ramirez et. al. 2012). It is important that faculty are fully aware of this fact so they can provide guidance to encourage students to make their theses openly accessible.

Those who argue that making theses openly available on the Web will lead to greater piracy of ideas and actual text tend not to examine the alternative of having the print theses readily available for borrowing on library shelves, with fewer tools and reviewers available to identify plagiarism. Open access enables full text indexing of theses that not only makes them more discoverable but also provides a basis for detecting plagiarism using online tools designed for this purpose. More "eyes" on the work can also curtail any claims to ideas presented in the work that others may present as their own after the thesis has been published.

The benefits of open access have been widely documented and are addressed in other documents in this collection.[5] At the highest level and applicable to ETDs, open access provides increased visibility for the ETD authors, their advisors, their funders and their institutions, increased citations resulting in greater impact (Eysenbach 2006), and prevention of duplicate effort in conducting research that has already been done (SPARC 2013). Institutions benefit by having the research they support made more visible to prospective students, faculty, and research collaborators, increasing the likelihood of attracting people and funding opportunities to their programs that align with the institution's strengths. These benefits may often out-weigh the reasons for restricting access to ETDs and should be given serious consideration when making decisions about the availability of ETDs.

2.4 Requesting Restricted Access

The policies surrounding who makes the decision about whether or not a thesis should be restricted will vary. Some institutions allow students to make that decision, and some require the thesis advisor to

[5] To find articles, presentations, webcasts and other open access materials, see the Scholarly Publishing and Academic Resources Coalition (SPARC) publications at http://www.arl.org/sparc/publications/index.shtml.

make the decision. Some require that the student and advisor make the decision together, while others have an institution-wide policy on embargoes for all ETDs. There can be a multi-level process for approving an embargo, similar to Rice University's policy where the dean of the student's academic unit must request the restriction with the recommendation of the thesis advisor, then it must be approved by the Dean of Graduate and Postdoctoral Studies (Rice University 2013). Other institutions may automatically grant embargoes simply on the basis of a student's request. The documentation and explanations required for requesting an embargo also vary widely.

The practice with print theses was often to hold them for a semester or a year, and then to send the manuscripts for processing. When the bound copies were returned to the library, the theses had already undergone a *de facto* embargo of sorts since they were not publicly available for up to eighteen months from the time they were first submitted. When transitioning to ETDs, some institutions have decided to impose a one-year embargo on all theses to mimic their availability as bound print works. As noted above, funders may also have requirements for restricted access that the students and institutions are required to acknowledge. In these situations, the request for restricted access should be accompanied by the sponsor's guidelines that clearly state the reasons for and terms of the ETD embargo.

Conflicts between institutional open access mandates for ETDs and sponsor policies requiring that a thesis be embargoed for some period of time will need to be resolved prior to final submission of the thesis. Sponsored research offices can play a role in negotiating with funders who want to restrict access to ETDs by letting them know of the institution's open access policy and encouraging the sponsors to minimize the restrictions they are placing on the research.

2.5 Enforcing Access Restrictions

ETDs are often managed in a repository and the repository manager may be responsible for enforcing any embargoes that have been placed on the theses. Some institutions, however, keep embargoed or other access restricted theses in the office of graduate studies or the college/school the student is associated with before releasing it to the repository. Depending on the sophistication of the repository and the workflow for submitting the ETDs, the document can be handled in multiple ways to ensure that an embargo is not compromised. When the decision is made by the departments, schools, or graduate studies offices to hold the embargoed theses and not submit them to the repository to be managed, there is a risk of not properly managing the ETD to ensure that the integrity of the document is guaranteed through long-term archiving and preservation practices that are followed with repository-deposited ETDs. It may also requires manual checks on a regular basis to identify theses whose embargo period has passed so that they can be added to the ETD repository.

Many repositories now manage embargoed ETDs with metadata or other markers that indicate there is an embargo. The repository software will automatically release the theses to make them available when the embargo period has passed. Many libraries are now responsible for managing ETD repositories, thus it is important that they know which, if any, theses have embargoes placed on them. It is also critical that the embargoed ETDs undergo the same preservation practices as the available ETDs to maintain their integrity.

2.6 Releasing Restricted Access ETDs

The stakeholders for a restricted access ETD should be aware of when an embargo is ending so that, if institutional policy permits, they may make a request that the thesis be embargoed for a longer period of time. The policies regarding renewal of ETD embargoes vary widely, as do the responsibilities for notifying the stakeholders that the embargo is ending. Embargoed thesis stakeholders can include the ETD author, faculty advisors, sponsors, academic departments, graduate student offices, compliance offices, and publishers. Institutions that permit an embargo to be extended must set clear policies on which stakeholders can request embargo extensions.

Institutions that allow embargoes to be renewed may choose to provide an advanced notice that the ETD will be released in a given number of weeks/months, while others place the responsibility on the stakeholders to stay abreast of the embargo timing. If institutions agree to provide advanced warning that the embargo is going to expire, they will need to maintain the contact information for the stakeholders so that they will receive the notice; this is not generally included in the metadata record, so it will usually be maintained in a system external to the ETD repository. Informing the stakeholders can be done by a reliable means of communication (e.g. email to a reliable email address or a letter to a physical address) and the stakeholders should be given a reasonable amount of time to respond. Whatever the policy is regarding notification of the end of an embargo, it is imperative that the stakeholders be informed of their responsibilities and the policies of the institution prior to a thesis being embargoed.

2.7 Summary

This guideline has identified a number of access restrictions and embargoes that can be applied to ETDs. The benefits of open access should be considered prior to applying restrictions to an ETD's availability. While some will argue that the students and their advisors should have the freedom to restrict access whenever they want (Hawkins, Kimball, and Ives 2013), the institutions and funders of the research have a vested interest in the scholarship that has been produced and will be the long-term stewards of the information. They will often want that scholarship to be disseminated to demonstrate the work that the institution and/or funders have supported. There are, however, occasions when a restriction does become necessary. Institutions must define consistent policies for applying any access restrictions and practices related to ETDs. Myths regarding publication prohibitions for openly available ETDs should be acknowledged as such, and any blanket policies to restrict access to support publication should be critically reviewed. Considerable variation exists in the practice of applying access restrictions to ETDs. As this medium becomes more prevalent, the best practices in this area will help to shape greater consistencies within and across institutions in establishing policies and procedures that provide authors with the greatest benefits while safeguarding any issues that could lead to harm. There is an increasing emphasis by funding agencies on sharing information that supports the trend towards increasing open access for ETDs. As these theses enjoy more citations than non-open access theses, there is more motivation on the part of students to ensure that access to their thesis is as open and free as possible.

Bibliography

Clement, Gail. 2012. "Graduate Student Manifesto." http://sites.tdl.org/fuse/?paged=2.

Department Of State. The Office of Electronic Information, Bureau of Public Affairs. 2011. "Overview of US Export Control System." March 8. http://www.state.gov/strategictrade/overview/.

Duke University Graduate School. 2013. "Duke Graduate School: Availability of Your Electronic Dissertation." *Duke University Graduate School*. Accessed January 3. http://gradschool.duke.edu/academics/theses/availability.php.

Education, UK Council for Graduate, Tina Barnes, Martin Moyle, Josh Brown, and Kathy Sadler. 2012. *Electronic Doctoral Theses in the UK: A Sector-wide Survey into Policies, Practice and Barriers to Open Access*. UK Council for Graduate Education. http://discovery.ucl.ac.uk/1339905/.

Eysenbach, Gunther. 2006. "Citation Advantage of Open Access Articles." *PLoS Biol* 4 (5) (May 16): e157. doi:10.1371/journal.pbio.0040157.

Graduate School, Virginia Tech. 2013. "Proprietary or Classified Theses and Dissertations." *Graduate Catalog – Virginia Tech Graduate School*. http://graduateschool.vt.edu/graduate_catalog/policies.htm?policy=002d14432c654287012c6542e3799999.

Hawkins, Ann R., Miles A. Kimball, and Maura Ives. 2013. "Mandatory Open Access Publishing for Electronic Theses and Dissertations: Ethics and Enthusiasm." *The Journal of Academic Librarianship* 39 (1) (January): 32–60. doi:10.1016/j.acalib.2012.12.003.

Kenney, A. R., R. Entlich, P. B. Hirtle, N. Y. McGovern, and E. L. Buckley. 2006. "E-Journal Archiving Metes and Bounds." https://clir.org/pubs/reports/pub138/pub138.pdf.

"LISTSERV 16.0 – ETD-L Archives – October 2012." 2012. *ETD-L Archives October 2012*. October. http://listserv.vt.edu/cgi-bin/wa?A1=ind1210&L=ETD-L - 9.

McCutcheon, A.M. 2010. "Impact of Publishers' Policy on Electronic Thesis and Dissertation (ETD) Distribution Options Within the United States". Ohio University. http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1273584209.

McMillan, Gail, Marisa L. Ramirez, Joan Dalton, Max Read, and Nancy H. Seamans. 2011. "An Investigation of ETDs as Prior Publications: Findings from the 2011 NDLTD Publishers' Survey." In Cape Town, South Africa. http://hdl.handle.net/10919/11338.

MIT Libraries. 2012. "About MIT Theses in DSpace@MIT: Document Services: MIT Libraries." *MIT Libraries Document Services*. February 23. http://libraries.mit.edu/docs/about-theses/.

National Institutes of Health (NIH). 2009. "Frequently Asked Questions – Public Access." *National Institutes of Health Public Access*. December 28. http://publicaccess.nih.gov/FAQ.htm - 758.

NDLTD. 2010. "ETD Terms and Definitions." http://www.ndltd.org/resources/Definition_of_ETD_Terms_6_10_2010_NDLTD.pdf.

OCLC. 2012. "WorldCat.org: The World's Largest Library Catalog." *WorldCat*. Accessed September 28. http://www.worldcat.org/.

Olson, John S. 2012. "Thesis Availability."

ProQuest LLC. 2012. "Home Page – ProQuest Dissertations & Theses (PQDT) – ProQuest." *ProQuest Dissertations & Theses (PQDT)*. http://search.proquest.com.ezproxy.rice.edu/pqdtft?accountid=7064.

"Questions and Answers – USPTO- USPTO." 2003. *United States Patent and Trademark Office: USPTO - General Questions*. August 14. http://www.uspto.gov/main/faq/index.html.

Ramirez, M. L., J. T. Dalton, G. McMillan, M. Read, and N. H. Seamans. 2012. "Do Open Access Electronic Theses and Dissertations Diminish Publishing Opportunities in the Social Sciences and Humanities?" *College & Research Libraries*. http://crl.acrl.org/content/early/2012/04/05/crl-356.short.

Rice University. 2013. "Candidacy, Oral Examinations and Thesis." *General Announcements 2012-2013*. April 22. http://ga.rice.edu/Home.aspx?id=123.

Seamans, Nancy H. 2003. "Electronic Theses and Dissertations as Prior Publications: What the Editors Say." *Library Hi Tech* 21 (1): 56–61. doi:10.1108/07378830310467409.

SPARC. 2013. "How Open Access Benefits Researchers (SPARC)." *SPARC*. http://www.arl.org/sparc/students/researcherbenefits.shtml.

United States Electronic Thesis and Dissertation Association. 2012. "USETDA » ETD Terms and Definitions." *USETDA: ETD Terms and Definitions*. http://www.usetda.org/?page_id=72 - E.

University of Kansas. 2011. "KU Policy| Embargo Policy for Theses and Dissertations." *The University of Kansas Policy Library*. October 20. https://documents.ku.edu/policies/Graduate_Studies/Embargo_Policy.htm.

University of Southampton. 2012. "ROARMAP: Registry of OpenAccess Repositories Mandatory Archiving Policies." *ROARMAP*. http://roarmap.eprints.org/.

USPTO. 2013. "First Inventor to File (FAQs)." *USPTO.GOV: The United Stares Patent and Trademark Office*. April 2. http://www.uspto.gov/aia_implementation/faqs_first_inventor.jsp.

3 Briefing on Copyright and Fair Use Issues in ETDs

Patricia Hswe (Penn State University)

Topics Covered

- Explanation of legal framework of copyright and fair use.
- Guidance on intellectual property rights education for ETD stakeholders.
- Implications of providing access to ETDs containing fair use materials.
- Issues of intellectual property rights when retrospectively digitizing ETDs.
- Effects of intellectual property rights on publishing portions of or depositing ETDs.

3.1 Introduction

Electronic theses and dissertations (ETDs) capture the research efforts of undergraduate and graduate students, many of whom will go on to pursue careers in which publications play a role in professional advancement. Universities and colleges have a responsibility to provide the best possible guidance on students' intellectual property rights – in particular, copyright and fair use. All parties involved in providing, supporting, and managing an ETD service should be apprised of the range of issues represented by copyright and fair use practices. Campus entities - such as the graduate school, the departments and programs it supports, the library, and the research administration office (e.g., the Office of the Vice-President for Research) – have their own stakes in an ETD service. It is in the interest of each of these stakeholders to ensure the preservation of and continued access to the scholarly record, provide copyright protection for research results shared in an ETD, and conduct workshops on ETD copyright and fair use as part of outreach to students and even to faculty.

The goal of this briefing document is to offer a variety of perspectives on copyright and fair use in the context of ETD service provision and management. It reviews copyright and fair use from the student author perspective (i.e., the fact that the student – as *author* of a thesis or dissertation – holds the copyright for it, or shares copyright with the institution accepting the ETD) and from the student user perspective (i.e., the fact that the student *uses* copyrighted material for integration in a thesis or dissertation). Besides students, the audience for this document includes ETD administrators, librarians, research administrators, faculty, and scholarly publishers. It advises on roles and responsibilities for communication of and training in copyright and fair use, both from an ETD service viewpoint and from the broader perspective of the academic institution housing the service, since research administration guidelines and policies impact ETD publishing and dissemination. In addition, it explores copyright issues stemming from the aggregation and delivery of ETDs by vendors such as ProQuest, particularly in this e-book age, and reports on current thinking about copyright in the context of retrospective ETD scanning projects. This briefing should lay a foundation for understanding the basics of these topics when administering an ETD service. It has some overlap with the *Guide to Access Levels and Embargoes of ETDs*. It also explicitly references *Metadata for ETD Lifecycle Management*; *Guidelines for Implementing ETD Programs – Roles and Responsibilities*; *Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs*; and *Managing the Lifecycle of ETDs: Curatorial Decisions and Practices*.

3.2 Definition and Overview of Copyright and Fair Use in an ETD Context

A baseline understanding of what copyright and fair use are, informed by a programmatic or curricular approach to literacy in these issues, should be de rigueur for students once they enter higher education. The reality, however, is that many students encounter these subjects for the first time only when immersed in researching and writing tasks for a thesis or dissertation. Students often include excerpts of copyrighted material in ETDs, incorporated perhaps to buttress an argument, or to display an image or other resource that is referenced, or to give further details for context. For these and other reasons related to inclusion and use of copyrighted content, students on the cusp of doing research for their theses and dissertations should understand what copyright and fair use mean. Libraries increasingly have personnel, such as copyright librarians, or copyright coordinators, who provide outreach and training in this area. Guidance on copyright and fair use, as well as on how to carry out an education program addressing these topics, is growing (Graveline 2011; Harper 2007). This knowledge not only will serve students well in an ETD context (e.g., educating them on open access issues), but also will inform their future scholarly publishing activities.

3.2.1 The Basics of Copyright Law

The US Copyright Office keeps current a circular on the basics of copyright, such as who is permitted to claim copyright; what types of work have copyright protection and what types do not; length of time copyright endures; and more. Copyright law was developed to protect "original works of authorship," whether published, or unpublished (US Copyright Office 2012). Works that are copyrighted may not be sold, or – in the case of works in the fine arts and performing arts – displayed or performed in public without permissions clearance. When a work is copyrighted, it is considered illegal to infringe on the rights procured by copyright law for the owner of the copyright. However, there are legal exceptions made for copyright accountability. Probably the best-know exception is the doctrine of fair use. This exception is discussed below." Another consideration students should be aware of is the option to incorporate materials that are in the public domain – i.e., materials that have fallen out of copyright, or are permitted for open, freely available use by the content creator.

Students who know they will be integrating copyrighted content that does not fall within the bounds of fair use (defined below) need to exercise due diligence, which means seeking permission from copyright holders. Graduate schools, as the administrative entity for handling ETD deposits, should collaborate with librarians who have knowledge of copyright and fair use issues on instruction and training sessions for students on how to contact copyright holders and draft requests for permission to use such copyrighted content.

3.2.2 Copyright Registration for ETD authors

As authors of ETDs, students hold the copyright to their theses and dissertations (or share copyright with an institution, depending on the latter's policies), which are considered original, authored works as well as their intellectual property. Institutions – such as Washington University, Cornell, and Oregon Health and Science University – suggest that students insert a copyright notice (©) in their ETDs. Others, such as American University, encourage insertion of a copyright statement. Most institutions provide guidance on how to assert copyright (whether through the symbol or a statement) in an ETD but *do not prescribe* any particular approach. ETD authors maintain copyright unless they transfer it – an intention that must be conveyed in written form. Since authorship of an ETD renders copyright ownership immediate, it is *not necessary* to register one's ETD formally for copyright. Registration is an option that typically incurs a fee. Authors of ETDs may register copyright directly with the US Copyright Office, or they may register it through ProQuest/UMI. Registration evidences formal ownership of copyright and thus, in the case of a thesis or dissertation, proof of authorship. Some institutions, such as Catholic University of America, recommend that students writing ETDs register formally for copyright.

3.2.3 International Copyright Law and US ETDs

Another issue related to copyright and ETDs is whether students' theses and dissertations are, as a matter of course, protected from copyright violations in other countries. While there is not an international copyright law, there are international copyright conventions in which the US is a member and that offer protection for US authors. These include the Berne Convention for the Protection of Literary and Artistic Works and the Universal Copyright Convention (UCC). Signatories of the Berne Convention enjoy what is called the "national treatment": works originating in a Berne Convention member country are accorded the same protective measures in each of the other Berne member countries that the latter allows for the works of its own nationals (Berne Convention 1979). This protection does not require formal registration or other formalities of Berne Convention members – it is automatic. A key factor to note is that sound recordings are not protected under the Berne Convention; this would be an issue for ETDs consisting of audio files, either in their entirety or in part. The UCC allows that "any formality under a national law can be satisfied by the use of a notice of copyright in the form and position specified by the UCC" (US Copyright Office 2012).

It is probably the rare ETD service at US institutions of higher education that encourages its student depositors to include a copyright statement asserting protection under the Berne Convention. (Indeed, a cursory review of literature on the topic of international copyright and ETDs yields no documentation on how many US institutions with ETD services provide this advice to their students.) US copyright law still applies in instances where ETD authors are international students, since they are submitting their theses and dissertations as part of the requirements of degree programs at US institutions. At the same time, because the Berne Convention is a recognized *international* copyright standard or agreement, it may be in the best interest of students, whether of US or other citizenship, to indicate that their theses and dissertations are protected from infringement under Berne as well as under the copyright law of their home countries. There is precedent in thesis and dissertation services at universities abroad for expressing protection under such a combination of international copyright convention and a nation's copyright law. One example of compliance under the Berne Convention comes from the University of Dar es Salaam; students submitting theses and dissertations (which are not necessarily electronic) sign a declaration that the thesis or dissertation is their original work, and underneath their signature appears the following statement:

> *The thesis is copyrighted material protected under the Berne Convention, the Copyright Act 1999 and other international and national enactments, in that behalf, on intellectual property. It may not be reproduced by any means, in full or in part, except for short extracts in fair dealing, for research or private study, critical scholarly review or discourse with an acknowledgement, without the written permission of the Directorate of Postgraduate Studies, on behalf of the author and the University of Dar es Salaam (Kiondo 2004).*

Whether beginning a new ETD service or auditing the practices and services of an existing one, the question of what to assert in a copyright statement for an ETD is one of policy (if not also of philosophy), requiring a decision on the part of the institution at the earliest possible point, since agreements such as terms of use and terms of service come foremost in the ETD deposit process.

3.2.4 The Basics of the Fair Use Doctrine

Besides knowing about copyright from the standpoint of an author, students writing ETDs should also be aware of copyright from the perspective of a *user* of copyrighted content. In addition to securing permission to use copyrighted materials in an ETD, students should be briefed that they have other options: to use public domain content or to apply the doctrine of fair use if incorporating copyrighted content. Apprising students of what defines public domain should be a part of any guidance on copyright and fair use in ETDs.[1]

Increasingly, scholarly communication librarians and copyright specialists based in academic libraries are providing such guidance by conducting workshops, creating informative web pages, and contributing to the drafting of guidelines to feature at institutions' ETD information websites. Good examples of library-based, advisory resources on copyright and fair use that are comprehensive (though not presented in the context of ETD preparation and submission) include the following: Columbia University Libraries/Information Services ("Copyright Advisory Office"); Emory Libraries ("Copyright and Publishing,"); Stanford University Libraries ("Copyright and Fair Use"); UCLA Library ("UCLA Copyright Policies"); and University of Minnesota Libraries ("Copyright Information and Resources").

Moreover, ETD service managers need to understand fair use in the context of hosting and making accessible theses and dissertations with third-party content. Fair use prevents copyright law from transgressing First Amendment rights, thereby offering a balance and flexibility to ensure continuation of expression unconstrained (Association of Research Libraries 2012). The fair use doctrine consists of four factors that should be weighed in determining whether the use of a copyrighted work qualifies as infringement or not. The US Copyright Office (2012) lists these factors as the following:

[1] One particularly useful resource is Peter Hirtle's "Copyright Term and the Public Domain in the United States" (http://copyright.cornell.edu/resources/publicdomain.cfm). Another online tool is one called "Is It Protected by Copyright?" (http://librarycopyright.net/resources/digitalslider/), produced by Michael Brewer and the American Library Association Office of Information Technology Policy.

a. The purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes.
b. The nature of the copyrighted work.
c. The amount and substantiality of the portion used in relation to the copyrighted work as a whole.
d. The effect of the use upon the potential market for, or value of, the copyrighted work.

In the UK, “fair use” is known as “fair dealing,” which may be claimed for the following types of content: “Research and Private Study,” “Criticism/Review,” and “News/Reporting.”

There is no prescription in fair use for *how much* of another’s work may be incorporated without appearing exploitative. Such setting of definite parameters could also be prohibitive and counteractive, going against the balance that fair use is meant to afford with copyright law. The doctrine walks a line between tractability and constraint. Yet, as the *Code of Best Practices for Fair Use* states, “Fair use is a user’s right,” (Association of Research Libraries 2012, 6). Such right will be as robust as the frequency of its exercise.

It has long been a practice of scholars to quote from or reproduce in part the work of another, reflecting an exchange of ideas that has potential to beget new ones. It is also been the purview of the creative and fine arts to use another’s work as a subtext for one’s own new work, whether parody, satire, or dramatic reinterpretation. For these reasons, an understanding of fair use early in the ETD research and writing process benefits students greatly. (It may also help prevent incidents of plagiarism, whereby students pass off another’s content as their own. See *3.2.5.3*.) In addition, ETD service managers should be aware of any guidelines or policies their institutions provide regarding appropriate use of copyrighted materials. In the absence of such guidance then, advice on fair use may need to be addressed on a case-by-case basis. Two helpful resources for assessing fair use are the Fair Use Evaluator (Brewer et al. 2008) and the Fair Use Checklist (Columbia University Center for Digital Research and Scholarship 2011).

3.2.4.1 The Association of Research Libraries’ “Code of Best Practices in Fair Use for Academic and Research Libraries”: On Libraries and ETDs

One perspective of the fair use doctrine examined in the Association of Research Libraries’ *Code of Best Practices for Academic and Research Libraries* is that of libraries with institutional repositories (IRs) housing ETDs. In the report, the Association of Research Libraries (ARL) asserts that it is incumbent on libraries, as stewards of such scholarship in digital format, to retain the integrity of ETDs, keeping any copyrighted content contained therein intact, rather than to demand permissions for or deletions of that content. Further, the *Code of Best Practices* recommends that, in cases where ETDs are hosted and maintained through a vendor application, libraries should require that vendors honor the rights of fair use applied by ETD authors. The fair use principle for the IR scenario is stated in the *Code of Best Practices* as follows: “It is fair use for a library to receive materials for its institutional repository, and make deposited works publicly available in unredacted form, including items that contain copyrighted material that is included on the basis of fair use,” (Association of Research Libraries 2012, 23).

The ARL report also notes current gaps in properly carrying out best practices for fair use. These include the need for a tool to make it uncomplicated for copyright holders to lodge complaints about use of their content in an IR and for libraries to respond to such complaints. According to the report, libraries and their home institutions should educate ETD-depositing authors about not only fair use but also the correct ways to attribute the inclusion of copyrighted material in an ETD, and about the tendency for fair use practices to be dependent on context – i.e., "fair use within the academy may not be fair when a work is more broadly distributed," (Association of Research Libraries 2012, 24). In addition, the *Code of Practices* argues that the case for fair use will be strengthened when institutions have a well-articulated policy about appropriate integration of third-party content, such as quotations and illustrations, in ETDs and other types of scholarly products. The report also suggests that libraries offer guidance on an individual basis regarding how to use copyrighted content in scholarly publications.

The sections that follow address various factors to consider in appreciating copyright and fair use in an ETD service context. These include an institution's research administration policies for intellectual property rights; who should create guidance on copyright and fair use, and for whom; what the implications for copyright are when digitizing theses and dissertations submitted before 1978 (works published without notice prior to 1978 are in the public domain); and how commercial publishers and vendors such as ProQuest impact our practices and our understanding of copyright and fair use.

3.2.5 Intellectual Property Rights, Sponsored Research, and Student Work

Since the advice that an ETD service provides on copyright and fair use depends in part on its local context, an understanding of the research administration policies and guidelines for intellectual property rights (e.g., patent and copyright) at one's institution marks an important prerequisite for such guidance.

Some theses and dissertations are based on research funded by a grant award, or supported through industry partnerships. Because of this fact, and because ETDs qualify as student work, it becomes imperative for an ETD service to investigate (and keep current on) institutional policies and guidelines related to intellectual property rights for research produced by faculty and students. In cases where ETDs will largely be the result of sponsored research, some institutions, such as Virginia Tech, require a review of that research as part of thesis or dissertation planning – or even a "pre-publication review" for determining publication restrictions, "including acceptable separation of restricted findings into an embargoed document," (Virginia Tech Graduate School 2012). In such cases, graduate schools urge students to apprise them early on if students anticipate restrictions to the research incorporated in a thesis or dissertation. An office of sponsored research or sponsored programs typically administers guidelines and policies that address this and similar issues.

In addition, ETD service managers may wish to consult the research administration guidelines and policies regarding any protections that their institutions have in place for students in the event that claims of copyright infringement or fair use violations are brought against them. This marks another scenario for which it behooves ETD service managers to confer with the office of sponsored programs on development of proper guidance for students, or to confirm that such guidance exists and is up to date.

3.2.5.1 Sponsored Research

When a thesis or dissertation incorporates sponsored research, a few questions to take into account include the following:

- What rights does the institution have to student work? When does research belong to the institution, and when does it not?
- In what circumstances can a student restrict access to a thesis or dissertation? When is an embargo appropriate, and for how long?
- Is there an institutional policy on research data management?

The rights of an institution to student work often depend on a variety of factors, including – but not limited to – the status of the student (undergraduate or graduate rank) and the context of the work (e.g., course requirement, or part of a funded research project). If a thesis or dissertation contains research relevant to a patent being filed, then a student is likely to place an embargo on the thesis or dissertation, delaying its public access. Another reason for restricting access is that occasionally grant-funded research must be reviewed prior to its distribution through scholarship (including ETDs), or that data are still in use on a project and cannot yet be made public. For more about embargoes and other concerns associated with access, please consult *Guide to Access Levels and Embargoes of ETDs*.

3.2.5.2 Research Data

In addition, some institutions, such as Johns Hopkins University, the University of Tennessee, and the University of Wisconsin-Madison, have formal policy statements concerning management of research data, in which parameters for retention and ownership of data are explained. Additional factors to weigh in this context are mandates from grant-funding agencies, such as the National Science Foundation (NSF) and the National Endowment for the Humanities (NEH), for the inclusion of data management plans with grant-proposal applications. These mandates are intended to facilitate the public's long-term access to research data that has been funded with taxpayer dollars. In light of these recent requirements, an ETD service manager may want to consult the office of sponsored programs on how best to advise students who are writing theses and dissertations based on research data generated from an NSF-funded project, for example; faculty serving as primary investigators (PIs) for such projects may need to be consulted as well, since occasionally the scholarship of theses and dissertations is mentioned as one of many venues through which data from the project will be disseminated. Liaison librarians who teach courses to both undergraduates and graduate students on the literature of a subject specialty, such as on chemical literature or biological literature, could work with ETD service managers to determine efficient paths for outreach and education regarding research data management in the context of writing ETDs.

3.2.5.3 Plagiarism

The topic of plagiarism tends to arise when copyright and fair use are addressed. While copyright seems closely related to plagiarism, copyright is steeped in the legality and economics of protecting an author's rights, whereas plagiarism has to do with a failure to cite resources supporting one's research (Vanacker 2011). Acting responsibly in the conduct of research means maintaining the integrity of that research, including giving proper attribution when and where credit is due. This does not mean, however, that accordance of attribution is equivalent to copyright compliance. Compliance occurs when permission to use copyrighted material is granted by the copyright holder to the user; or when the use of copyrighted content falls under fair use; or when the content being used is in the public domain.

Students writing theses and dissertations may ask whether ETDs are more vulnerable to plagiarism than print theses and dissertations, because ETDs are – by and large – rendered immediately accessible (University of Pittsburgh 2007). A typical answer to this question is that anything that is published risks being plagiarized. Yet, there are measures students can take to deter or prevent copying or extraction of content from their ETDs. Software applications (such as TurnItIn[2]) can be used to assist in automating, to some extent, detection of plagiarism. Since students who write ETDs are advised by faculty regularly, it is also incumbent on thesis and dissertation advisers to read *carefully* the drafts their students submit periodically for review.

3.2.6 Providing Guidance about ETD Copyright and Fair Use: Who, for Whom, When?

Figuring out which collaborative parties should be involved in creating and providing guidelines on copyright and fair use for students writing ETDs is a key initial step. At the same time, it is important to note that students are not the only audience for such guidelines – that is, some of the very entities needed for collaboration may require guidance on copyright and fair use, too.

3.2.6.1 Who Provides Guidance: Forming an "ETD Collaborative" on Campus

As implied by the foregoing, ETD services are informed by cross-departmental partnerships on a campus, including – but not limited to – administrative bodies such as the graduate school, undergraduate program, and office of sponsored programs; faculty, who make up the teaching, research, and learning arm of an institution; librarians, whose specializations support the teaching, research, and learning activities of faculty and students; and the institution's general counsel. Each has a slightly different stake in ensuring the best possible guidance for students engaged in ETD preparation. From the perspective of an office of sponsored programs, the primary goal might be to encourage integrity in conducting research, while from the perspective of an office of general counsel, compliance with institutional and legal guidelines and policies may be the focus. For a more detailed picture into the roles and responsibilities that support an ETD service, please refer to *Guidelines for Implementing ETD Programs – Roles and Responsibilities*.

Together, these entities make up an "ETD collaborative," which should work toward uniform, consolidated guidance for addressing copyright and fair use in ETDs. Being "on the same page" is critical. Students benefit greatly from a comprehensive, integrated overview: it saves them the time of looking all over an institution's website, and even beyond, for the relevant pieces of information, and it sets the example that a host of factors should be considered in order to understand and address copyright and fair use adequately in scholarship.

[2] See http://turnitin.com/.

3.2.6.2 Guidance for Whom: The Internal and External Audiences

The audience for the guidance provided by a campus-based ETD collaborative is, foremost, internal, consisting primarily of students at the institution that is home to the ETD service. In addition, it should include those with direct contact with students and/or their scholarship, such as faculty, liaison librarians, and institutional repository (IR) managers or scholarly communications librarians. An ideal would be to target *all* students – whether undergraduate, master's, or doctoral – from the start of their academic careers, well before they enter the ETD stage, rather than to wait to require ETD-writing students to attend workshops on these issues, or point them to the relevant resources in the nick of ETD-filing time. It would also be ideal for ETD service managers to work closely with faculty, librarians, and IR managers – the latter are typically well-versed in scholarly communication issues such as copyright, fair use, and open access. Education and outreach for these constituencies carries advantages at once preventative and proactive. For example, faculty who co-author publications with students need to be apprised of copyright practices, including transfer of copyright, in order to advise students properly of their rights in collaborative authorship situations and in order to avoid inadvertently signing away students' copyright (Clement 2012). Liaison librarians should stay current on copyright and fair use as a matter of course in order to convey suitable advice to students as part of their research services.[3]

Besides an internal audience, external audiences are likely to find ETD copyright and fair use guidance of benefit. Scholarly publishing organizations, including university presses, may wish to find out what policies and practices for intellectual property rights were at play for theses and dissertations they are interested in accepting for publication, whether in part or in their entirety. The literature on ETDs and implications for scholarly publishing may be helpful in this regard (Ramirez et al. 2012; McMillan 2001). Other universities and colleges that have decided to launch ETD programs may be in search of examples against which to benchmark the services they are beginning to model. For these reasons, ETD service managers should consider documenting copyright and fair use guidance openly, making it publicly and easily accessible.[4]

3.2.6.3 Guidance for When: From the Start

Knowledge of copyright and fair use serves students invaluably from the start of their undergraduate and graduate careers. Attuning all students to copyright and fair use in the context of their scholarship *before* they arrive at the ETD phase has implications for efficiency in their research and writing and for the integrity of ETD and other scholarly content. A background in these issues also bodes well for their potential future as faculty members or as other types of researchers: as part of publishing their

[3] See the section above, titled "The Basics of Fair Use Doctrine" for a list of guides and resources that can help keep librarians posted on these topics. In addition, the Berkman Center for Internet and Society at Harvard has created an tutorial called, "Copyright for Librarians" (http://cyber.law.harvard.edu/copyrightforlibrarians/Main_Page). Another excellent method for keeping up to date on copyright and fair use is to read blogs such as "Copyright Librarian" (http://blog.lib.umn.edu/copyrightlibn/), by Nancy Sims, Copyright Program Librarian at University of Minnesota Libraries, and "Scholarly Communications @ Duke" (http://blogs.library.duke.edu/scholcomm/), which Kevin Smith, Scholarly Communications Officer at Duke University, writes.

[4] Good examples of ETD services providing such documentation include, but are not limited to, Duke University (http://gradschool.duke.edu/academics/theses/copyright.php) and North Carolina State University (http://www.ncsu.edu/grad/etd/docs/etd-guide.pdf).

research, they will be asked, for example, to secure permission for using any copyrighted images or illustrations, or to submit a signed author agreement as required by a publisher (publishers as a matter of course pass this responsibility on to their authors). First-hand exposure to copyright and fair use issues, including the deposit agreement(s) students are obliged to understand, can amount to a formative authorship experience. Yet, the guidance should not be limited to the relevance it has to ETD writing and submission activities. That is, there is an opportunity for institutions, via an ETD service and the collaborating research library, to be strategically proactive in teaching students about authorship experiences they may have beyond their degree programs. Thus, ETD service managers should consider working with their institution's libraries, including liaison librarians and scholarly communications personnel, and with their institution's legal counsel to develop recommendations. These recommendations could address the following: how to review a publisher's contract properly and negotiate for one's author rights; how to publish in open-access journals; and how to deposit data sets into open disciplinary repositories for broad discoverability, access, use, and reuse.

3.2.7 Distribution of ETDs via an Institutional Repository

Often, the indexing and accessibility of ETDs are managed in the context of an IR. This section surveys copyright and fair use in an IR/ETD service context. It takes into account concepts such as distribution agreements, "take down" policies, and author agreements. The section also considers possible research scenarios that ETD service managers may wish to present to students as a way to kick-start thinking about the benefits of sharing and making openly available their research.

3.2.7.1 Agreements and Licenses

IRs typically promote open access of scholarly materials, making ETDs a logical content fit for them. In addition, students writing ETDs own the copyright to their work and retain that copyright following the deposit of the thesis or dissertation into an IR. The copyright status of items such as ETDs can be conveyed via a rights statement in the metadata. Generally, IRs do not exert any rights to the content deposited in them. Most IRs require that depositors consent to a non-exclusive distribution license – i.e., permission, granted by ETD authors, for the IR to archive, make publicly accessible, and manage the ETDs. The purposes of archiving and dissemination go hand in hand (e.g., there is no dissemination without ongoing archiving), although ETD depositors might not understand this duality at first. However, making the case for both, equally and strongly, exposes students further to the challenges of digital preservation and archiving. Reviewing such licenses is also not unlike reviewing an author's agreement; therefore, gaining experience reading and understanding license agreements prepares students for reviewing author agreements and publishing contracts in the future. Students should be reminded of Section 106A of the Copyright Act, which asserts rights of attribution and integrity in authorship. In addition, some IRs are supporting Creative Commons licenses,[5] which depositors may choose from at the time of ingest and metadata entry. These license options could also be presented to students depositing ETDs. Obviously, if the thesis or dissertation contains research that cannot yet be shared, such as when a student is filing a patent, then an embargo option may be warranted.

[5] See http://creativecommons.org/licenses/.

3.2.7.2 Discovery Implications for Open Access ETDs

While many institutions provide straightforward guidance regarding information such as the above, few of them describe – in the same guidance – scenarios about what it means to make one's thesis or dissertation openly accessible. These scenarios include ETD titles appearing in Google or Bing search results or via other digital pathways. ETD service managers who oversee deposits of thesis and dissertations may wish to explore what measures their IRs are taking to increase search engine optimization (SEO), which focuses on how to lend a website or web page more impact and thus increase traffic to it; it is about "affecting the visibility of a website or a web page in a search engine's 'natural' or un-paid ('organic') search results," (Wikipedia 2013). Some students do not wish for their works to be found easily, while others champion such broad access. Equally important is informing students what could happen if copyright holders, perhaps discovering via a search in Google or Bing that their content has been used in a thesis or dissertation, claim copyright infringement or violation of fair use. In the event of such situations, ETD service managers should know what their institutions' policies are; what role the IR plays in such actions; the relevance of the Digital Millennium Copyright Act (DMCA); and what options are available to students whose ETDs are involved. Many ETD programs and IRs have "take-down" policies, whereby the thesis or dissertation is removed from public access because of a copyright or fair use dispute. More information on ETDs and open access may be found in the *Guide to Access Levels and Embargoes of ETDs*.

3.2.7.3 Documenting ETD Usage and Impact

Related to scenarios of ETD discovery are scenarios of ETD use, which aid in measuring the impact ETDs have on scholarship. Many IRs that distribute ETDs are able to provide usage statistics, including the overall number of downloads since ingest, or the number of downloads in a month, and how users are coming to their work, or the item record for it (e.g., via a Google search, or a link to it from an online citation manager such as Mendeley). Informing students of these types of usage statistics can help them see the benefit of making their work accessible worldwide. In addition, ETD service managers may wish to investigate new uses of ETDs based on current events in scholarly publishing, such as a recent agreement between a researcher and Elsevier in early 2012 to data mine runs of journals published by Elsevier (Howard 2012).

Given that some ETD programs have been in existence for a decade or more, and thus have a substantial number of theses and dissertations, data mining and text mining requests could emerge in the near future for ETDs in certain subject areas. There is evidence in the literature that data mining should not reflect copyright infringement, since, for example, "most scientific data are facts that are not covered by copyright except to the extent that an author has exercised minimal creativity in the selection or arrangement of data," (Carroll 2011). In anticipation of researchers making requests to mine the data and text of ETDs, managers of IRs and ETD services may wish to formulate guidelines and a process for responding to such requests, as well as point researchers to recommended software tools for getting out data and text.

3.2.8 Copyright and Fair Use in Retrospective Reformatting of Theses and Dissertations

Some institutions, whether they have an ETD service already in place or not, have initiated retrospective scanning of theses and dissertations that are in print only, as well as those that were submitted in print and then microfilmed by UMI/ProQuest. Practitioners have begun to address questions of copyright in a conversion context, and the section below touches upon some of these.

3.2.8.1 Why Retrospectively Digitize?

In the last few years, some libraries have started digitizing legacy, or historic, theses and dissertations. The reasons for doing so include the following:

- For both librarians and users, the myriad formats (often in print and as a UMI/ProQuest microfilm copy) and locations (e.g., archives, special collections, or subject libraries) of these materials impede easy access to them. Digitizing this content for online access facilitates improved search and discovery (Shreeves and Teper 2012).
- Legacy theses and dissertations are often made available via inter-library loan, which has inherent access constraints – such as a lending period of only a few weeks. Many of these items are also in a fragile state; instead of lending out the print volumes, libraries often scan them as PDFs and mount them online. The online availability gives researchers unrestricted access, unlike inter-library loan, and serves a preservation purpose. At the same time, these are "one-off" situations that can disrupt staff workflows. Thus, digitizing and thus making digitally available a body of historic ETD content can eliminate such disruptions.
- Through scanning these dissertations and theses, libraries may encourage open access of scholarship as well as strengthen the academic reputation of their institutions (Martyniak 2008).
- Historic theses and dissertations are considered grey literature, and some libraries are committed to broadening access to such materials as part of strategic collection development activities.

A key issue to consider in retrospective conversion of historic theses and dissertations from print to digital format is their copyright status, which the library literature has started to address. As most of the articles discussed below attest, the tactics to take in determining copyright status include consulting with legal counsel at one's institution to see where it stands on this issue; negotiating with commercial entities that make such content available at a price so that institutions can have some control over it for the purpose of broader access; and working with groups such as alumni associations, colleges, departments, and graduate schools to establish contact with thesis and dissertation authors for securing their permission to digitize, and render available online, their past scholarship.

3.2.8.2 Brief Literature Survey

Clement and Levine (2011) investigated whether pre-1978 dissertations, in accordance with the 1909 Copyright Act, count as publications or not, a dependency in determining their copyright status. They found that "For copyright purposes, these were acts of publication with the same legal effect as dissemination through presses, publishers, and societiesm" (Clement and Levine 2011, 825). They suggest that collection managers should investigate the copyright status for dissertations deposited in libraries, including those microfilmed by UMI/ProQuest, between 1909 and 1978; if there is no copyright notice, then the thesis or dissertation is likely in the public domain. Moreover, some of these dissertations might have fallen out of copyright, if they were not renewed after 28 years for the same length of time (Clement and Levine 2011, 826).

Shreeves and Teper (2012) recount their experience of a pilot project to digitize historic theses and dissertations at the University of Illinois at Urbana-Champaign (UIUC), in which their first priority was a thorough understanding of rights issues. For this they consulted University Counsel, which approved online access to these materials provided it was limited to the UIUC community; an additional proviso was that the library give copyright holders – the authors of the theses and dissertations – the options of removing their digitized content, or of opening it up for worldwide "open access" exposure (Shreeves and Teper 2012, 533). Furthermore, while University Counsel also approved the conversion of theses and dissertations in paper to digital format, they were hesitant when it came to digitizing these materials in microfilm, since the microfilmed copies were essentially the property of ProQuest. ProQuest subsequently proposed a plan in which UIUC would cover the cost of digitizing the microfilmed theses and dissertations and permit their access through ProQuest's service. Negotiations over a three-year period ultimately landed in UIUC's favor: ProQuest agreed to let the university provide restricted access to the digitized content via its institutional repository, as well as to allow open access with permission from thesis and dissertation authors. Another hurdle in this pilot, however, was the process of securing such permissions from the authors. Again, ProQuest and UIUC worked on a solution suitable to both parties that streamlined the permission process. Thus, an important factor in considering retrospective conversion of these legacy items is the possible need to negotiate with vendors like ProQuest for control over such local content.

The ETD-L mailing list[6] has also been a venue for discussion of questions related to copyright and fair use in the context of retrospective conversion of theses and dissertations, as well as of suggestions for solutions. Since the list has subscribers from all over the world, it is not uncommon to get glimpses of ETD service experiences abroad. For example, the service manager for the Electronic Theses Online System, or EThOS, in the U.K. posted a response to a question from a US university about which institutions are doing digitization of legacy theses and dissertations *without* first seeking permission of the authors. The EThOS service manager referred the list to a "Frequently Asked Questions" (FAQ) section on the EThOS site, in which appears the answer to the question, "How has the issue of obtaining retrospective permissions for digitization been dealt with?" (EThOS Administration 2012). The answers that EThOS gives include the following: 1) EThOS argues that it is a cost-recovery operation and does not profit from digitization of theses and dissertations, whereas the authors of them, as well as the institution where these materials were deposited, enjoy enhanced discoverability and recognition of the intellectual content; 2) it is unrealistic to believe permission from all thesis and dissertation authors can be obtained; and 3) there are "take-down" policies in place in the event that authors reject such easy availability of their scholarship.

Finally, both Martyniak (2008) and Shreeves and Teper (2012) note the important role that alumni associations, colleges, and departments can play in contacting, or locating, authors for their permission to digitize and make accessible their theses and dissertations. Each of these articles describes the convoluted processes that permissions work can involve; on the other hand, the latter does mention the creation of an online form that authors would fill out to confirm ownership of copyright for a thesis or

[6] See http://listserv.vt.edu/cgi-bin/wa?A0=ETD-L.

dissertation as well as whether they favor or not making their content openly accessible (Shreeves and Teper 2012, 534).

3.2.9 ETDs, Publishers, and Publishing

As the previous section implies, vendors such as ProQuest and other commercial publishers have a vested interest in managing and promoting ETDs. With the proliferation of the e-book format, institutions need to keep abreast of issues relevant to graduate students who have completed and submitted ETDs and thus own the copyright to them. There is great potential for ETD service managers to work with librarians, copyright specialists, faculty, and publishers on assembling better guidance to equip students for publishing their scholarly work.

As stated at the beginning of this document, the experience of writing theses and dissertations gives students a sense of what it is like to prepare a scholarly work for publication. There are format and submission standards to which ETDs adhere, just as there exist standards for the preparation of scholarly manuscripts; there are committees consisting of faculty members who vet and advise on the substance and quality of ETDs, not unlike what an editorial team does; and students are asked to consent to agreements (such as non-exclusive distribution agreements). The framework for preparing and depositing ETDs is analogous to, and portends, various stages of scholarly publishing. Thus, additional guidance for ETD service managers, librarians, and even faculty with scholarly publishing experience to provide for graduate students writing theses and dissertations could address questions of publication in the context of ETDs: Can an ETD include a chapter that has been published as an article? What should students know about reviewing publication agreements? How should students be advised about ProQuest options? What impact, if any, do ETDs have on students' future publication prospects?

3.2.9.1 Prior Publication and ETDs

It is not uncommon for students to publish a portion of their theses or dissertations, such as a chapter, as an article prior to submission. Policies and practices surrounding this issue may depend in part, however, on the local institutional context of the ETD service – i.e., whether or not review committees will allow students to integrate previously published material in their theses and dissertations. More important is whether the student has transferred copyright to the publisher, or retained the right to use the material. Students should read thoroughly their publishing agreements to make certain they understand what is allowed. Accordingly, if there is anything in such agreements that students take issue with, then they should be encouraged to negotiate the agreement with publishers. A student who has transferred copyright to a publisher and would like to reuse the published content for a thesis or dissertation will need to contact the publisher for permission to do so. Some ETD services, such as at Duke University, advise that rather than integrating a chapter that has already become an article, students discuss the research behind the article in a distinctive way. This approach avoids violation of copyright law, which safeguards the *expression* of an idea, not the idea itself (Duke University Graduate School 2012). Finally, students should be apprised of resources, such as Sherpa RoMEO,[7] to aid them in

[7] See http://www.sherpa.ac.uk/romeo/.

figuring out various publisher policies for depositing previously published content into a repository service.

3.2.9.2 ProQuest and ETDs

Many ETD services offer students the option of making their theses and dissertations also available via ProQuest Open Publishing PLUS.[8] A key advantage of paying ProQuest to host one's thesis or dissertation is expanded means of discoverability: the ProQuest service can, if students choose, expose theses and dissertations to search engines; moreover, the ProQuest Theses and Dissertations database has been known to receive a couple hundred million searches a year (Hadro 2010).

With ProQuest, students still retain and own the copyright to their theses and dissertations, but the service has the non-exclusive right to distribute the ETDs. Unlike with the traditional publishing option, whereby students receive royalties from the sale of their theses or dissertations, with the Open Access Publishing PLUS option, students are permitting free, worldwide access to their work, potentially in any format, including as an e-book. Within the constraints of not being legal practitioners, ETD services would do well to review, in concert with librarians and copyright specialists, the publishing agreement furnished by ProQuest, in order to give the best possible guidance to students.

3.2.9.3 ETDs and E-Books

The recent phenomenon of people finding the ETDs they submitted as students on sale as e-books suggests that close review of the agreement with ProQuest is crucial, particularly where students are given the choice of distribution of their ETDs via "third-party selling" (Torres 2012), or third-party sales. For, even though a student holds copyright of her thesis or dissertation, if she chooses third-party sales as an additional way of distributing her scholarship, then vendors such as Amazon and Barnes and Noble are within their rights to sell that thesis or dissertation as an e-book, with profits going to them but not to the student. Another concern that ETDs published as e-books raises is whether the e-book format negatively affects future publication of the thesis or dissertation as original scholarship as journal articles and monographs (Smith 2012). It is early days yet for the ETD as e-book phenomenon, but, generally, since ETDs being published as e-books are not revised or put through a peer-review process, then this concern is effectively moot; e-books are simply another format – just as the microfilm of a thesis or dissertation that is then scanned as a PDF and bound in cloth as a book encompasses just another format.

3.2.9.4 Publishing Potential of ETDs

A common concern among students, particularly those with aspirations for tenure-track academic positions that rely heavily on the publication of original scholarship, is that making an ETD widely accessible hampers their chances at publishing it in print. This notion is actually a misguided one, and getting the facts and trends straight can impact the success of outreach and educational activities intended to promote the benefits of ETDs (McMillan 2001). A 2011 survey of journal editors and university press directors in the social sciences, arts, and humanities found that, for the most part, a manuscript's prior status as an electronic thesis or dissertation did not affect its publication potential,

[8] See http://www.proquest.com/en-US/products/dissertations/epoa.shtml.

mainly because publishers expect – and require – ETDs accepted for publication to be substantially revised beforehand (Ramirez et al. 2012). Moreover, ETDs do not undergo external peer review, which is a process required by journals and university presses; for most journal editors this fact still makes ETDs, or parts of them (e.g., for articles), eligible for submission as original scholarship (Howard 2012). Enhanced access to a students' research through its status as electronic theses or dissertations may even *lead* to publishing opportunities. The idea that making a thesis or dissertation open access is a deterrent to formal publication as articles or as a monograph is important to rectify for faculty, too. Faculty advise students on their theses and dissertations and need to be current on today's scholarly publishing trends.

3.3 Summary

ETD service managers, as well as librarians, faculty, and institutional administration, should recognize there are opportunity costs if key issues in copyright and fair use are not presented as thoroughly as possible for all stakeholders involved in an ETD service. We do students an important, relevant service in giving them the tools they need to make informed decisions about copyright, fair use, and author rights. Such a service extends to the institution at large as well, in that it can lessen the likelihood of legal action. By coordinating efforts and thus displaying a more centralized, unified front in the understanding of copyright and fair use practices, institutions can reduce or prevent confusion among faculty and students; possibly shape more efficient development of research policies and guidelines; and position themselves to think more strategically about future research and the future uses of research.

Bibliography

Association of Research Libraries. 2012. “Association of Research Libraries’ Code of Best Practices for Academic and Research Libraries.” http://www.arl.org/pp/ppcopyright/codefairuse/code/index.shtml.

Berkman Center for Internet and Society at Harvard University. “Copyright for Librarians.” http://cyber.law.harvard.edu/copyrightforlibrarians/Main_Page.

Brewer, Michael, and ALA Office for Information Technology. 2008. “Fair Use Evaluator.” http://librarycopyright.net/resources/fairuse/.

---. 2012. “Is it protected by copyright?” http://librarycopyright.net/resources/digitalslider/.

Carroll, Michael W. 2011. “Why Full Open Access Matters.” *PLoS* 9, no. 3. doi:10.1371/journal.pbio.1001210.

Clement, Gail, and Melissa Levine. 2011. “Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach.” *Portal: Libraries and the Academy* 11, no. 3: 813–829. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v011/11.3.clement.html.

Clement, Gail. 2012. "Copyright Hot Topics." Paper presented at the Texas ETD Association Conference. Denton, Texas, February 23.

Columbia University Library/Information Services. “Copyright Advisory Office.” http://copyright.columbia.edu/copyright/.

Columbia University Libraries/Information Services. 2011. "Fair Use Checklist." http://copyright.columbia.edu/copyright/fair-use/fair-use-checklist/.

Duke University Graduate School. "Academics: ETD Publishing Concerns." http://gradschool.duke.edu/academics/theses/publish.php.

"Electronic Thesis Online System (EThOS) Toolkit." http://ethostoolkit.cranfield.ac.uk/tiki-view_faq.php?faqId=3-q56.

Emory Libraries. “Copyright and Publishing.” http://web.library.emory.edu/copyright-and-publishing.

"ETD-L@LIST.VT.EDU." http://listserv.vt.edu/cgi-bin/wa?A0=ETD-L.

The Graduate School, North Carolina State University. 2011. “Electronic Thesis and Dissertation Guide.” http://www.ncsu.edu/grad/etd/docs/etd-guide.pdf.

Graveline, Jeffrey D. 2011. “Launching a Successful Copyright Education Program.” *College & Undergraduate Libraries* 18, no. 1: 92–96.

Harper, Georgia. 2007. "Copyright Crash Course." http://copyright.lib.utexas.edu/index.html.

Hirtle, Peter, Emily Hudson, and Andrew Kenyon. 2009. “Copyright and Cultural Institutions: Guidelines for Digitization for US Libraries, Archives, and Museums”. http://papers.ssrn.com/abstract=1495365.

Hirtle, Peter. 2012. “Copyright Term and the Public Domain in the United States.” http://www.ncsu.edu/grad/etd/docs/etd-guide.pdf.

Howard, Jennifer. 2012. "Elsevier Experiments with Allowing Text Mining of Its Journals.” *The Chronicle of Higher Education*, May 6. http://chronicle.com/article/Hot-Type-Elsevier-Experiments/131789/.

Johns Hopkins University, 2008. "Policy on Access and Retention of Research Data and Materials." http://jhuresearch.jhu.edu/Data_Management_Policy.pdf.

Kiondo, Elizabeth. 2004 "Historical Practice in Managing Theses and Dissertations at African Universities and University Libraries." DATAD Workshop on Intellectual Property, Governance, Dissemination and Funding Strategies, Accra, Ghana, February 19-20. http://www2.aau.org/datad/reports/2004workshop/kiondo.pdf.

Martyniak, Cathleen L. 2008. “Scanning Our Scholarship: The University of Florida Retrospective Dissertation Scanning Project.” *Microform & Imaging Review* 37, no. 3. doi:10.1515/mfir.2008.013.

McMillan, Gail. 2001. "FEATURES-CONFERENCE CIRCUIT-Do ETDs deter publishers?-Does Web availability count as prior publication?" *College and Research Libraries News* 62, no. 6: 620-622.

Perry, Mark, and Paula Callan. 2009. “Legal Protocols and Practices for Managing Copyright in Electronic Theses.” http://papers.ssrn.com/abstract=1495722.

Ramírez, Marisa L., Joan T. Dalton, Gail McMillan, Max Read, and Nancy H. Seamans. 2012. “Do Open Access Electronic Theses and Dissertations Diminish Publishing Opportunities in the Social Sciences and Humanities?” *College & Research Libraries*. http://crl.acrl.org/content/early/2012/04/05/crl-356.

Shreeves, Sarah L., and Thomas H. Teper. 2012. “Looking Backwards Asserting Control over Historic Dissertations.” *College & Research Libraries News*73, no. 9: 532–535. http://crln.acrl.org/content/73/9/532.

Smith, Kevin. 2012. "Dissertations for Sale, or Scaring the Children, Part 2." http://blogs.library.duke.edu/scholcomm/2012/07/07/dissertations-for-sale-or-scaring-the-children-part-2/

Stanford University Libraries. “Copyright and Fair Use.” http://fairuse.stanford.edu.

Torres, Manuel R. 2012. “Dissertation for Sale: A Cautionary Tale.” *The Chronicle of Higher Education*, June 24. http://chronicle.com/article/Dissertation-for-Sale-A/132401/?cid=wb&utm_source=wb&utm_medium=en.

US Copyright Office. 2012. "Copyright Basics." http://www.copyright.gov/circs/circ01.pdf.

US Copyright Office. 2012. "Fair Use." http://www.copyright.gov/fls/fl102.html.

US Copyright Office. 2012. "Rights of Certain Authors to Attribution and Integrity." http://www.copyright.gov/title17/92chap1.html.

University of California at Los Angeles. “UCLA Copyright Policies.” http://www.library.ucla.edu/copyright/ucla-copyright-policies.

United Nations Educational, Scientific, and Cultural Organizations (UNESCO). 1971. "Universal Copyright Convention." http://portal.unesco.org/en/ev.php-URL_ID=15241&URL_DO=DO_TOPIC&URL_SECTION=201.html.

University of Minnesota Libraries. “Copyright Information and Resources.” https://www.lib.umn.edu/copyright/.

University of Nottingham, SHERPA Services. 2011. "Publisher Copyright Policies and Self-Archiving." http://www.sherpa.ac.uk/romeo/index.php?fIDnum=|&mode=simple&la=en.

University of Pittsburgh, 2007. "ETD FAQ." http://www.pitt.edu/~graduate/etd/faq.html.

University of Tennessee, Office of Research. "Research Data Policy." http://research.utk.edu/forms_docs/policy_research-data.pdf.

University of Wisconsin-Madison. 2011. “Policy on Data Stewardship, Access, and Retention.” http://www.grad.wisc.edu/research/policyrp/rpac/documents/PolicyDataStewardship.pdf.

Vanacker, Bastiaan. 2011. “Returning Students’ Right to Access, Choice and Notice: A Proposed Code of Ethics for Instructors Using Turnitin.” *Ethics and Information Technology* 13, no. 4: 327–338.

Virginia Tech Graduate School, 2012. "Graduate Catalog 2012-2013: Policies, Procedures, Academic Programs." http://graduateschool.vt.edu/graduate_catalog/policies.htm?policy=002d14432c654287012c6542e3720025.

Wikipedia, the Free Encyclopedia. 2013. “Digital Millennium Copyright Act.” http://en.wikipedia.org/wiki/Digital_Millennium_Copyright_Act.

Wikipedia, the Free Encyclopedia. 2012. “Fair Dealing.” http://en.wikipedia.org/w/index.php?title=Fair_dealing&oldid=527201614.

Wikipedia, the Free Encyclopedia. 2012 “Public Domain.” http://en.wikipedia.org/w/index.php?title=Public_domain&oldid=528272701.

Wikipedia, the Free Encyclopedia. 2013. “Search Engine Optimization.” http://en.wikipedia.org/wiki/Search_engine_optimization.

World Intellectual Property Organization. 1979. “Berne Convention for the Protection of Literary and Artistic Works.” http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html.

4 Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs

Yan Han (University of Arizona)

Topics Covered

- Benefits of collecting ETD usage metrics.
- Examples of usage metrics published by several American ETD programs.
- Methods to gather both quantitative and qualitative data.
- Explanations of common web statistics to gather.
- Methods to analyze return on investment (ROI) and open access benefits.

4.1 Introduction

Individuals (authors, faculty, and graduate students), institutional structures (libraries and graduate colleges), and the scholarly community in general (users) play different roles in the electronic thesis and dissertation (ETD) curation lifecycle. These roles include content generation, delivery, access, use and reuse, and preservation. In order to assess and understand the outcomes of ETD programs, these stakeholders increasingly seek information regarding title and/or collection usage, impact, and user satisfaction rates. In order to demonstrate the value of ETD programs, stakeholders that provide content generation, delivery, and access to end users (usually libraries and graduate colleges), must collect and produce usage metrics.

Producing usage metrics involves the measurement, collection, analysis, and reporting of usage for a title or a collection. Usage data is the most important data collected to measure outcomes, such as return on investments (ROI), effectiveness of delivery methods, and the characteristics of a collection (e.g. subject-oriented collections have concentrated users but high visits). As a result, it is critical to collect accurate usage metrics and to do so while respecting users' privacy.

4.1.1 Benefits of Usage Metrics for an ETD Program

Libraries and graduate colleges can use quantitative data such as ETD web visits and web-visit-demographics to understand users, collections, and impact. In combination with qualitative data such as survey and focus groups, they can do further data mining and analysis to find out more information such as ROI and user satisfaction. Specific case studies can be found in *4.2* below.

An institution, such as the graduate college or library, needs to understand the use of its resources and the outcomes of its investments. Institutional administration is interested in the big picture such as the overall impact of a collection, the institution and/or department rankings, the costs, and the ROI. The graduate college might be interested in knowing more specific details regarding its scholarly outputs and its graduate students' experiences with the ETD program. The library may be most interested in the users of its resources and their experience with the repository software and search mechanisms that enable them to access the collections.

As it seeks to assess its ETD program, an institution needs to make the connections between (a) resources and services in supports of its institutional goals; and (b) how the resources and services are used, by whom, and their impact. These measurements and statistics are critical justifications for an institution to receive financial and administrative support for advancing its scholarly outputs and providing access to information resources and their related services. Areas that an institution needs to evaluate include, but are not limited to the following:

- Satisfaction of graduate students from the institutional view (e.g. evaluation of services offered, new and better services for future students).
- Effectiveness of the overall ETD program (e.g. submission and delivery) and ROI.
- Usefulness of the ETD collection and its impact.
- Effectiveness and efficiency of delivery methods.

As the end users of the ETD collections, researchers are more likely to be interested in understanding the impact of a particular title and/or a field. They see the usage metrics in a unique light, including:

- Measuring the impact and importance of a title.
- Evaluating the impact of their ETDs (e.g. citation usage, potential collaborators).
- Assessing the impact of certain research fields within an institution (e.g. citation and usage data to compare fields/programs offered by peer institutions).
- Evaluating the satisfaction of graduate students from individual view (e.g. theses/dissertation publishing and copyright services).

4.2 Evaluation of Electronic Resources: Methods and Issues

Both quantitative and qualitative approaches should be taken in evaluating digital resources (such as ETDs) as a resource and as a service. As a resource, assessment focuses on how the *collection* was used. As a service, assessment focuses on how *users* interact with digital resources (Franklin, Kyrillidou, and Plum 2009). As with other digital materials, changes in content delivery and users' behaviors have driven fundamental changes in assessment methods and tools.

4.2.1 Examples of Measuring Usage in US Institutions

Many institutions collect usage statistics, but might not collect them consistently or be compliant with best practices. Since the web has been evolving over time, usage statistics and best practices also change regularly. However, basic concepts and models have been consistent. It is strongly recommended to collect usage statistics with respect to resources available and to keep the statistics archived to demonstrate the outcomes.

Some institutions provide certain usage statistics for public view, which adds value to the collection and repository because it helps general users to assess the uses and impact. Below are just a few noteworthy examples.

4.2.1.1 Virginia Tech Usage Statistics

Virginia Tech is known as the first university to require ETD submissions going back to 1997. They publish the following metrics:

- Usage statistics such as the number of HTML pages and PDFs accessed and unique visitors for its ETD collection since 1997. The quantitative data gathered and synthesized by Virginia Tech provide important statistics for the university and ETD community to see the impact and the growth over the past 16 years.[1]
- Annual surveys of ETD authors and users. For example, in the 2006-2007 ETD author survey there are 14 closed-ended and open-ended questions providing feedback regarding issues such as submission, workshop, preparation, file formats, and committee involvement.[2]

4.2.1.2 University of North Texas Usage Statistics

The University of North Texas (UNT) repository also provides its usage statistics for public view, including items added, usage per title, partner, and collection.[3] Alemneh assessed the ETD usage in the UNT Libraries and concluded that there are challenges and opportunities to providing access to digital resources (Alemneh 2011).

4.2.1.3 Texas A&M Graduate Student Survey

Texas A&M University's graduate college has been conducting a graduate students theses/dissertation survey as an important measure of students' overall university experience. To date it shows that the majority of graduate students were satisfied with their graduate study experience including ETD submission (Dromgoole 2012).

4.2.2 Overview of Evaluation of Library Resources: Methods and Issues

Libraries have a long history of evaluating and studying use of library resources and collections. In a traditional print library, librarians count various outputs such as collection size and collection usage data (e.g. circulation numbers), reference question numbers and types, numbers of inter-library loans, etc. It is a challenge to have a consistent and reliable way to collect these numbers due to variances in methods and sample sizes. These challenges are not new to the digital library. As noted by Glavin and Kent in 1977, library data and metrics were "too global in character and too imprecise in nature to serve as an adequate basis for reformulation of acquisitions policies. It is useless to tell the acquisitions librarian that half the monographs ordered will never be used, unless we can specify which 50 percent to avoid buying,“ (Glavin and Kent 1977, 2317-2320).

Regardless of the challenges inherent in data collection and analysis, libraries have used both quantitative and qualitative approaches to assess the performance of their digital library infrastructures. Some quantitative measures and tools include Counting Online Usage of Networked Electronic Resources (COUNTER), Standardized Usage Statistics Harvesting Initiative (SUSHI), and general web statistics such as ScholarlyStats. COUNTER provides consistent, credible and comparable usage from a

[1] See http://scholar.lib.vt.edu/theses/data/somefacts.html.
[2] See http://lumiere.lib.vt.edu/surveys/results/.

variety of vendors/publishers, while SUSHI is a standardized protocol for an electronic resource management system with COUNTER data. COUNTER and SUSHI are complementary initiatives designed to improve the reliability and usability of online usage statistics. These data are analyzed and generally used as cost-per-use data. Collecting meaningful cost data is not easy; however, one popular analysis is ScholarlyStats, which claims to report more than 400,000 e-journals. Though COUNTER and SUSHI are typically used for measuring performance of e-journals, the cost-per-use data these tools help to generate might be interesting to apply to ETDs to see the cost-per-use for ETD titles.

4.2.3 Quantitative Approaches

4.2.3.1 Collection Statistics

As the Internet fundamentally continues to change the way people communicate and share information, libraries see a profound increase in acquiring and serving networked digital resources rather than traditional materials. Digital resources become the de facto standard for information delivery. The Association of Research Libraries (ARL) has worked on new measures for the evaluation of electronic resources. Since 1961, the association has published *ARL* Statistics, "a series of annual publications that describe the collections, expenditures, staffing, and service activities for ARL member libraries." In terms of web statistics, ARL started project E-Metrics in 2000 to collect data about electronic resources and services. The E-Metrics project was carried out in three phases: a) Phase I was to gather inventory of ARL libraries and database vendor statistics; b) Phase II was to collect and analyze data; and c) Phase III was to propose measurement for electronic resources (Miller and Schmidt 2001). E-Metrics was designed to measure electronic information resources. The measures were designed to: a) be consistent with organizational missions, goals, and objectives; b) be integrated with an institution's program review; c) balance customer, stakeholder, and employee interests and needs; d) establish accountability; and e) include the collection and use of reliable and valid data.

The project studied a self-selected group of 24 libraries, and found that most libraries kept track of:

- Types of electronic materials.
- User measures (e.g. number of logins/visits and numbers of resources accessed).
- Types of users of electronic resources and services.
- Costs (e.g. cost per electronic document delivered and cost of database subscription)
- Other measures related with electronic resources and services such as survey, LibQUAL+, and focus group.

The project compared 12 major database vendors, showing that they collected general web statistics such as document types, sessions, visits, logins, and searches. Many vendors complied with the International Coalition of Library Consortia (ICOLC) guidelines drafted in 1998; however, practices range widely.

[3] UNT's ETD collection statistics can be viewed at http://digital.library.unt.edu/explore/collections/UNTETD/stats/.

Lack of consistent definitions, comparable measures, and standardized reporting methods topped the list. The E-Metrics project called for standardizing usage reports, sharing project information, and developing a set of core measures. A number of statistics and measures were recommended, including Patron Accessible Electronic Resources, Use of Networked Resources and Services, Expenditures of Networked Resources and Related Infrastructure, and Library Digitization Activities. The recommended statistics and measures were designed with library content and services in mind, which covers (a) technical infrastructure, (b) information content, (c) information services, (d) support, and (e) management (Shim et al. 2001).

ARL currently collects terms, including number of collections, number of items, size in GB, number of items accessed, and number of queries conducted (searches). The emphasis is collection-oriented and related to academic libraries' traditional measurement approaches such as such as those for collections, sizes and usage. There are notable issues with these statistics. For example, number of queries conducted is not included in the Digital (Web) Analytics Association (DAA) web analytics definitions and therefore web analytics tools such as Google Analytics do not report this measure. Library staff should devise a way to report this term in a standardized fashion.

4.2.3.2 Web Analytics

In the past, academic libraries tended to measure content and impact by collecting internal statistics, primarily collection-oriented (e.g. collection size and spending budget) and not user-oriented (e.g. users' behavior and users' experience). It has been a decade since ARL published its E-Metrics results. E-Metrics and performance indicators cannot fully provide libraries with users' perceptions and assessments of their services (Bertot and Davis 2004).

To face the changing information landscape, academic libraries should collect qualitative and quantitative measures to understand more about their end users and services. Business web sites such as retailers, advertising companies, and marketing companies focus on users' behaviors and experiences. Their goal is to collect, analyze, measure and report Internet data in order to understand and optimize web usage. Web analytics is not only a good tool for measuring web statistics, but also a rich data source for business intelligence and marketing research. Web analytics generally use two methods:

- Log file analysis: This method reads and interprets the log files recorded by a web server such as Apache or IIS. Common terms (see *Table 4-1*) can be recorded, and HTTP errors can be captured as well. Log file analysis is easy to do because web servers generate the raw data automatically. The main issue with log files is accuracy, resulting from a browser's cache capacity.
- Page tagging: To address pitfalls of log file analysis, tagging methods using JavaScript and/or an invisible image have been. Unlike log file analysis, page tagging also works for non-HTML web pages such as interactive Flash movies. It can be also used for companies who do not have access to their own web servers. Therefore, page tagging is widely used in web analytics.

After years of consolidation, web analytics terms tend to be consistent though there is still no national or international standard. Easy-to-use web analytics tools are available from different companies. The most popular one is Google Analytics (offered as both a free version and a fee-based premium version). Google Analytics makes it very easy to gather basic quantitative measures, including those suggested by E-Metrics project. In addition, Google Analytics provides a lot of user-oriented data, including demographics (e.g. country, city), behavior (e.g. new, returning users, frequency), technology (e.g. browsers, network), and mobile devices. These statistics can be, and often are, used for business intelligence and data mining for marketing, content delivery optimization, infrastructure and system improvement. More on how libraries can use such approaches for quantitatively assessing their ETD collections is included below.

4.2.3.3 Altmetrics

Altmetrics, a new metrics proposed in 2010, is an alternative to widely used *impact factor*. Its purpose is not limited to citation counts and is not limited to articles only. Altmetrics can be applied to journals, books, data sets, web pages, and others. It measures extended impacts such as views, discussed, downloads, mentions in social media. Like other metrics, controversy has been arisen as altmetrics can be self-cited, gamed and boosted in other ways.

Some publishers such as BioMed Central, Public Library of Science and Elsevier have started to provide altmetrics. The National Information Standards Organization (NISO) has been awarded a 2-year grant in June 2013, and is working on a project to study, propose and develop standard(s) and practices in altmetrics. Since altmetrics is an emerging way to measure impacts, it is recommended to keep an eye on the development of the standards and best practices.

4.2.4 Qualitative Approaches

Qualitative research involves studying and collecting a variety of empirical materials such as case studies and interviews, along with interactive and visual observations, all with the goal of identifying meaning to individuals. It is reasonable that each individual is different, and therefore the research has to study more than one interpretive practice (Denzin and Lincoln 2011). Employing multiple methods of qualitative research can offer better understanding of a research topic. Each method has its history, uses, context, and implementation. Two commonly used methods include surveys and focus groups.

4.2.4.1 Surveys

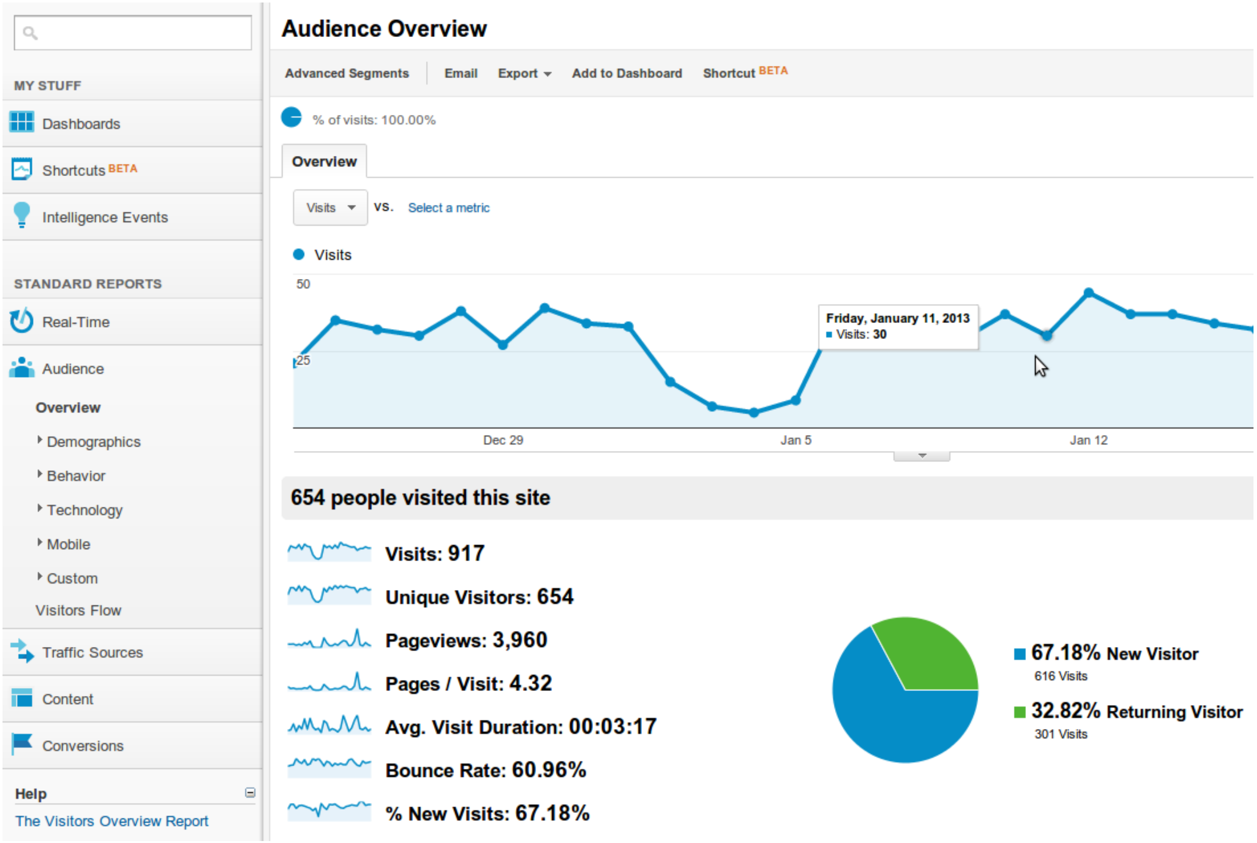
Surveys are used to study representatives from a population. For example, polls of public opinions are reported in the news media. Since it is based on a sample of the research population, the success is dependent on the degree of representation. This method has its advantages and pitfalls. The advantages include standardization, ease of management, cost-effectiveness (cost is low compared to focus groups), and efficiency for collecting information on a large population. One should be aware that the challenges are how to: (a) identify samples; (b) design, evaluate and adjust questions; and (c) reach out and contact those who are reluctant to respond. User surveys can be performed during ETD submission to understand graduate student experience including ETD submission. Within the research library domain, ARL maintains Measuring the Impact of Networked Electronic Services (MINES), an online survey service to collect data for the use of electronic resources. Kyrillidou, Plum, and Thompson have also presented a literature review of library web surveys and methodologies, and provided a set of methods to evaluate electronic services to better serve research, teaching and learning (2010, 159-183). A recommended guidance tool for surveying is *Survey Methodology* (ISBN 978-0470465462) by Groves et al. from the University of Michigan Survey Research Center.

4.2.4.2 Focus Groups

Focus groups involve a moderator facilitating a small group discussion on a topic. Advantages include valuable insight to data unlikely generated through personal interviews and observations, as well as opportunities to discuss the group's experiences. It was found that "focus groups often produce data that are seldom produced through individual interviewing and observation" (Kamberelis and Dimitriadis 2005). The disadvantages are the limitations of a one-time-study (unless repeated), and the risk that the focus group approach could collect biased data if the setting is not right. To learn more about conducting an effective focus group, we recommend the book *Focused Interview* by Robert K. Merton who is the inventor of the focus group methodology (1952).

4.2.5 Recommended Approaches

In general, an ETD collection from an institution is accessible through the institutional repository. The repository system market is dominated by a few such as DSpace and CONTENTdm. These systems either provide a way to integrate with Google Analytics or provide their own web statistics. Some institutions also set up third party web analytics software to collect detailed usage about users' behaviors.



654 people visited this site

Visits: 917

Unique Visitors: 654

Pageviews: 3,960

Pages / Visit: 4.32

Avg. Visit Duration: 00:03:17

Bounce Rate: 60.96%

% New Visits: 67.18%

■ 67.18% New Visitor
616 Visits

■ 32.82% Returning Visitor
301 Visits

Figure 4-1. Google Analytics screenshot

4.2.5.1 Minimal Level (Quantitative)

We recommend that institutions take an approach that meets its basic needs, as an institution may have resources and skills constraints. Assuming that the institution has an existing repository, the effort to collect usage metrics is straightforward and does not require many additional resources and skills. Web analytics tools such as the Google Analytics Standard version are free of charge and provide recommended web measurements. Some repository systems such as DSpace have Google Analytics code built in, and only need to have it enabled via a configuration file. Collecting statistics such as basic ARL measures is also relatively straightforward. Most repository software systems provide basic collection statistics such as the total number of items, and the total size of a collection. Using the above web statistics, one can report on the major terms defined in the ARL E-Metrics project. One item, number of queries conducted may need additional standardization if the repository software does not provide it. In this case, library staff may look at the Google Analytics "content" section for the "search" page (see *Figure 4-1*).

Resources and/or skills required (time, skills and resources are based on the author's 10 years of experience working with usage statistics):

- Setup (time, skills and resources): 0.5 to 1 hour to set up if the system includes modules for analytics code as DSpace does. However, If a repository system does not have this built in, it would most certainly require either using (1) adding analytics module, or (2) using third-party software and administrative access to web server logs. No IT skills or financial resources are needed.
- Maintenance (time, skills and resources): None (repository system/Google provided).
- Reporting (time, skills and resources): 0.5 to 1 hour on a recurring basis to output usage statistics using web interface.

4.2.5.2 Advanced Level (Quantitative and Qualitative)

For institutions that have additional resources and would like to know more about their collections, impact, and user experiences, we recommend the collection of more data through both quantitative and qualitative approaches. Quantitative and qualitative approaches require:

- Quantitative data: Minimal level (see above), plus other measures such as content analysis and user technology (e.g. browser, mobile devices). These measures may help design better delivery methods and user interfaces.
- Qualitative data: Surveys and/or focus groups. Surveys can be easily implemented in the curation lifecycle of ETDs, from submission to access. In order to produce reliable data and to avoid design pitfalls such as sample selection bias, one must understand the limitations and advantages of these methods. Two helpful starting resources could include: LibQUAL+, which is a web-based survey for libraries to solicit, track and response users' opinions of a service; and MINES for Libraries an online survey to collect data on the use of electronic resources and user demographics.

Term	Definition	Use	Note
Hit	A request for a file from the web server	The number of hits has been often cited for usage, but this could be misleading and over-estimates without understanding what can count as a hit.	A single web page can consisting multiple files (e.g. images, javascripts, stylesheets, PDF), each of which is counted as a hit when the page is accessed.
Page View (Pageview)	The number of times a page was viewed	Page view provides more realistic assessment.	A single page view may generate multiple hits.
Visit (Session)	An individual visits a website (usually consisting of one or more page views). If no action happens within a specified time period, the visit will be over.		Different providers use different methods for tracking. A typical specified time-out period is 30 minutes. More details see Web Analytics Definition.
Unique Visitor (Unique User)	A distinct person who visited a website. A visitor can make multiple visits.		Uniqueness is tracked by cookies (no login) and IP / user agent authentication.
Entry page	The first page in a visit.		
Visit Duration	Time spent in a visit		
Bounce Rate	A bounce occurs when a visitor only views a single page. Bounce rate = Single page visits / entry pages	Used to determine the effectiveness of an entry page. "a 50% bounce rate is average. If you surpass 60%, you should be concerned". For reference please see Inc.com [4]	
New Visitor	A visitor that has not made any previous visit during **a reporting period.**	"New visitor" is helpful in determining loyalty of a site, when comparing to "return visitor".	Each visitor should be counted as a NEW visitor **only once**; however, technology is not perfect.
Return(ing) Visitor (Repeat Visitor)	A visitor that has made previous visits during a reporting period	New visitors + Return visitors = Unique visitors	Some tools call "Repeat visitor" is the same as "Return visitor".

Table 4-1. Base terms to collect for measuring services

[4] http://www.inc.com/guides/2011/01/how-to-reduce-your-website-bounce-rate.html.

Setup, maintenance, and reporting (time, skills, and resources) will all vary when taking more advanced level approaches. For example, there may be fees for customizing web analytics tools if advanced web measurements needed, and certainly collecting qualitative data will take more resources and time. This work pays off in the quality of the data an institution gathers.

4.3 Collecting Web Statistics

4.3.1 Recommended Web Statistics

Web statistics data can provide invaluable information for ETD managers and administrators. It shows where users are from, what users visited, how long users stayed and how many titles users viewed. Interpretation of this information is very useful for institutional intelligence. Several web statistics tools have been widely used and data has been effectively gathered. As previously noted, Google Analytics is the most popular web analytics tool.

When reviewing web analytics, it should be noted that each tool/system might use different terms. *Visit*, *hit*, and *page view* might not necessarily mean the same thing across different systems. For example, Google Analytics uses "visit" where other tools use "session." Currently there is no standard or *de facto* reference for web statistics, although some web analytics associations such as DAA or committees are trying to create some guidelines. Libraries are encouraged to collect the following base terms in *Table 4-1* for measuring services, in addition to ARL E-Metrics statistics.

When analyzing web statistics, one should be aware that web crawlers (e.g. Googlebot, Bingbot, etc.) regularly visit and index websites, including repositories. Most analytics software such as Google Analytics use a simple strategy to validate non-web-crawlers visits, creating a 1x1 pixel clear-GIF image that is non-cacheable and not indexed by web crawlers. As a result, data from these analytics software including Google Analytics is fairly accurate (i.e., it does not include web-crawler visits in its data). If you run your own log analysis tool, you will need to evaluate your data and eliminate these web-crawler visits to ensure that your data is accurately representing use by researchers, not "noise" from other agents.

The terms in *Table 4-1* are recommended to collect for a minimal level set of web statistics. Important ratio statistics such as page/visit can be produced once the minimal set is collected.

4.4 Use of Statistics Data

One use of collection statistics is for the Ranking Web of World's Repositories.[5] This resource provides a list of mainly research-oriented repositories arranged according to a composite index of their web presence and impact. It uses indicators such as size, visibility, rich files and scholar content, and these indicators are combined to calculate the ranking. For example, arxiv.org is ranked the 2nd place in the world repository ranking with 18th in size, 1st in visibility, 1084th in rich files, and 2nd in scholar. Obviously increasing the number of titles and quality of titles can improve a repository ranking.

[5] See http://repositories.webometrics.info/.

With user measures and the cost model available, return on investment (ROI) figures can be calculated. Usually people talk about cost per title. Institutions can use these ROIs to compare published literature as one demonstration of their outcomes. Scant literature has been published to discuss the usage of grey literature, but we expect that more studies will be available in the future so that a better understanding of ROI can be reached.

Common ROIs suggested include:

- Cost per title: Annual costs related to one title. Linn (2009) states that "there are so few good examples of how librarians can use cost-benefit analysis." The well-known arXiv.org maintained by Cornell University listed per-article costs as $9 in 2006, $6 in 2009, and $8 in 2012; while the French HAL archive listed per article costs as $6.5 (€ 5) in 2008 (Schopfel and Boukacem-Zeghmour 2012).
- Cost per item request: Annual costs related to one item request. A case study (Piorun and Palmer 2008) suggested that the cost per item (ETD) request over a specified period of time was $1.90 (total project cost: $23,562; 17,555 downloads in 17 months, in comparison to 723 uses for print dissertations).
- Item requests per collection: Average access number per item in a collection. Two case studies show that ETDs are requested twice as much as articles and reports (Schopfel and Boukacem-Zeghmour 2012).
- Costs per user: Annual costs related to a user.
- Costs per author: Annual costs related to an author.

For more on ROIs for ETDs, see *Guide to ETD Program Planning and Cost Estimation*.

There can be issues in consistency of collecting usage data and even within COUNTER-compliant reports, there is evidence that the cost analysis is more difficult than it appears (Conyers and Dalton 2007). Combining quantitative data (web statistics) and qualitative data (surveys, focus groups) will help to analyze user behaviors and ROI.

It is possible to measure not just usage but also impact in the form of citations to articles in a repository. Software like Citebase does this, though it requires a significant body of interlinked content to produce meaningful metrics. For example, Citebase works well on the high-energy physics open access database, arXiv, because there is enough material in that database (over half a million articles) to permit meaningful measurement of citations between articles. As the open access corpus grows worldwide in institutional repositories, Citebase and other tools like it should be able to work effectively on a larger body of scholarly literature.

There is an increasing focus on the development of metrics that can measure aspects of research. In the past, the metric that has been in widespread use is the journal impact factor (JIF), which is the most well known metric in wide use by publishers, librarians and scholars. Developed by the Institute for Scientific Information (ISI), the JIF is used to measure the impact of individual journals, currently covering around 10,000 journals. As open access literature and repositories increase, there are concurrent increases in the need and ability to measure the impact of literature similar to the JIF. Open access global literature, indexes, and usage metrics will provide librarians and scholars around the world with the raw material to develop many new metrics for measuring and assessing research.

4.5 Open Access Model and Impact

When an institution decides to provide access to its ETD collection, it may choose between two main access models: open access (OA) and restricted access. Using usage statistics and citation analysis, many articles and studies have been published to analyze the impact of OA and non-OA literature. Overall, studies consistently demonstrate that OA provides easier access and higher impact over non-OA publishing models.

The following bullets provide a brief synopsis of some of the literature on this topic since 2001.

- Computer Scientists have found that OA articles have greater impact and more citations due to easier availability (Lawrence 2001).
- In 2004 Thomson ISI did a study of the OA journals in core scientific publications, and concluded that OA itself does not necessarily create more or fewer citations by comparing 191 OA journals and the 8509 non-OA journals indexed by ISI. The increase of readership does not necessarily increase a journal’s impact. Therefore, OA journals can have similar impact to other journals, and authors should not avoid publishing just because of the access model (Pringle 2004).
- Harnad and Brody made a further analysis on the Thomson ISI study by comparing the citation impact of OA vs. non-OA articles in the same journals. They draw a conclusion of OA having dramatic citation advantages. In 2006 they re-confirmed this conclusion (Harnad and Brody 2004; Harnard et al. 2006, 36-40).
- Antelman also found that OA articles do have a greater research impact in all four disciplines analyzed (Antelman 2004, 372-382).
- Eysenbach found that strong evidence shows OA articles are “more immediately recognized and cited by peers than non-OA articles published in the same journal” (Eysenbach 2006).
- Impact factor was tested again in 2010 to compare OA journals vs. non-OA journals; Giglia concluded that OA journals can compete with non-OA journals (Giglia 2010).
- Wacha and Wisner suggested that current research uses quantity, not quality, to assess value by studying 48 American repositories. They concluded that institutional repositories are not consistently collecting materials with high impact, and suggested further research for actual impact (Wacha and Wisner 2011).

Theses and dissertations are classified as “grey literature” in the scholarly community, which is not available through the usual systems of publication and distribution. In traditional libraries, grey literature is either not catalogued at all or is only partially cataloged, and therefore grey literature is not easy to find and access. OA promotes unrestricted online access to content, and under the OA publishing model, authors can share work broadly to reach a wider audience. Over the past decade, most institutions have set up their own institutional repositories, and some societies have also established subject repositories such as the renowned arXiv for math, physics, astronomy and computer science. Scholarly OA content has increased steadily in recent years, and OA ETDs account for the biggest collection of OA content. The nature of grey literature and the movement towards OA make theses and dissertations a perfect candidate for wider distribution and access. The experience reported (see below) shows that ETDs published under OA provide all kinds of benefits to scholarly communication, institutions, faculty and students. As a result, we strongly recommend publishing ETDs as part of your institution's OA repositories with the understanding that there may be multi-faceted embargo issues that will require clear policy (see *Guide to Access Levels and Embargoes of ETDs*).

4.5.1 OA ETD Benefits to Scholarly Communication, Institutions and Students

Scholarly communication is the beneficiary of OA. Free online availability makes access much easier, including direct links (no proxy), better indexing by search engines and easier integration with third-party services such as content integration. Vijayakumar and Vijayakumar (2007, 67-75) discussed the importance and dissemination of dissertations, surveyed the global ETD movement, and suggested that academics play a critical role in creating scholars and meeting challenges at national and international levels. Other OA implementations have been analyzed such as Korea Institute of Science and Technology ETD (Zhang and Lee 2001, 555-566), UK strategy and the Edinburgh Research Archive (MacColl 2002) (Jones et al. 2006), China Networked Digital Library of Theses and Dissertations (Jin 2004, 367-370), University of Waterloo and Theses Canada in Canada (Jewell, Oldfield, and Reeves 2006, 183-196), Ben-Gurion University in Israel (Asner and Polani 2008), KIIT University in India (Swain 2010), and North Carolina Universities (Early and Taber 2010).

4.5.1.1 Benefits to Institutions

Institutions are actors and beneficiaries of OA ETDs. Universities and colleges, big or small, public or private, carry out roles of teaching and research. To facilitate both teaching and research, publishing research and scholarship is one of the critical missions. By making ETDs open access, universities and colleges make ETD content easily accessible without restrictions and reach broader audiences.

- Benefits to teaching: OA ETDs add value to the student learning experience. Graduate students can easily find and access related research without traditional grey literature boundaries such as being difficult to find and not available online.
- Benefits to research: OA ETDs avoid lengthy traditional publishing timeframes, encourage uses, and increase visibility of university research. In return, web statistics and measures of collection and services (including ETDs) give fair indicators of a university's contribution. For example, Webometrics's repository ranking has been published since 2008, which is used to measure a repository's visibility and scientific impact, and ARL statistics measure size of content and visits.
- Benefits to discovery: OA ETDs allow indexing services from search engines and libraries. Exposing metadata and full-text documents makes them available to scholars.
- Cost savings: Libraries have already reported that OA ETD programs provide significant cost savings. By moving to OA ETDs, institutions remove traditional workflows for paper-based theses and dissertation, which means no more printing, binding, shelving and manual cataloging. (Jewell et al. 2006, 183-196).

4.5.1.2 Benefits to Graduate Students

Graduate students themselves are the contributors and the beneficiaries of OA. Benefits include:

- Better access to related research: The availability of ETDs enhances discovery and access to related research for graduate students and their advisors. Both primary and secondary data are available for ready access.
- Better presentation of research: One article suggested that ETDs have the advantage to present both primary data and secondary interpretation in accessible forms (Macduff 2009). There is a trend of increased use of multimedia formats and data sets coming with ETD submissions. It is reported that less than one third of survey respondents noted that non-text formats in ETDs was important (Lippincott and Lynch 2010).
- Enhancing graduate study experience: Ease of submission and publication enhance graduate students' experience. Council of Graduate Schools released a report in 2010, which suggested a few practices to enhance students' experiences including dissertation support (Council of Graduate School 2010). For example, a Texas A&M University graduate college survey shows that graduate students are happier with its ETD support in place (Dromgoole 2012).
- Cost savings: Graduate students save money by experiencing lower printing, submission and publishing fees when an OA ETD model is in place. University of Waterloo had a case study on demonstrating cost savings and remote submission for students (Jewell, Oldfield, and Reeves 2006, 183-196).

4.6 Summary

We strongly recommend collecting usage metrics including qualitative and quantitative data to demonstrate the value of ETD programs. At the very least, we recommend collecting a minimal level of usage data with tools such as Google Analytics. Institutions with additional resources are encouraged to take an advanced level approach for better understanding their users and impact of their ETD programs. A set of web statistics terms that are commonly used can become the core set for statistical analysis. Along with costs in mind, ROIs can also be analyzed and can be compared with available literature. These ROI figures and usage metrics can work together to demonstrate the efficiency and effectiveness of the ETD program.

Bibliography

Alemneh, Daniel Gelaw. 2011. “Assessing the Usage of Electronic Theses and Dissertations: An Overview of ETD Statistics and Metrics in the UNT Libraries” presented at the Texas ETD Association 2011 Conference, March 31, Arlington. http://digital.library.unt.edu/ark:/67531/metadc32969/.

Antelman, K. 2004. “Do Open-Access Articles Have a Greater Research Impact?” *College and Research Libraries* 66 (5) (September): 372–382. http://hdl.handle.net/10760/5463.

ARL. 2010. *E-Metrics: Measures for Electronic Resources*. http://www.arl.org/stats/initiatives/emetrics/index.shtml.

Asner, Haya, and Tsviya Polani. “Electronic Theses at Ben- Gurion University: Israel as Part of the Worldwide ETD Movement.” *Portal: Libraries and the Academy* 8 (2): 121–139. doi:10.1353/pla.2008.0021. http://muse.jhu.edu/journals/pla/summary/v008/8.2asner.html.

Bertot, J. C., and D. Davis. 2004. *Planning and Evaluating Library Networked Services and Resources*. Westport, CT: Libraries Unlimited.

Betty, Paul. 2009. “Assessing Homegrown Library Collections: Using Google Analytics to Track Use of Screencasts and Flash-Based Learning Objects.” *Journal of Electronic Resources Librarianship* 21 (1): 75–92.

Bradley, N. 2007. *Marketing Research. Tools and Techniques*. Oxford: Oxford University Press.

Clifton, Brian. 2010. *Advanced Web Metrics with Google Analytics*. 2nd ed. Sybex.

Conyers, A., and P. Dalton. “Electronic Resource Measurement: Linking Research to Practice.” *Library Quarterly* 77 (4): 463–470.

Council of Graduate Schools. 2010. *Ph.D. Completion and Attrition: Policies and Practices to Promote Student Success*. Washington DC: Council of Graduate Schools. http://www.phdcompletion.org/information/book4.asp.

Denzin, Norman K., and Yvonna S. Lincoln. 2011. *The SAGE Handbook of Qualitative Research*. 4th ed. Los Angeles: Sage Publications.

Dowling, Thomas, and Simon Bevan. 2010. “Online Usage Statistics for ETDs.” In Austin: UT Austin Libraries. https://conferences.tdl.org/index.php/utlibraries/etd2010/paper/view/76.

Dromgoole, Christine. 2012. “Check Yes or No: Analyzing Student Satisfaction with the Texas A&M University Thesis Office.” Poster presented at the TxETDA 2012 Conference, Danton. http://txetda.wordpress.com/etd-forum/2012-txetda-annual-conference/2012-conference-program/.

Early, Mary G., and Anne Marie Taber. 2010. "Evolving in Collaboration: Electronic Thesis and Dissertation Workflows in North Carolina." *Collaborative Librarianship* 2 (1). http://collaborativelibrarianship.org/index.php/jocl/article/viewArticle/56.

Eysenbach, G. 2006. "Citation Advantage of Open Access Articles." *PLoS Biology* 4 (5) (May). doi:10.1371/journal.pbio.0040157.

Fang, Wei. 2007. "Using Google Analytics for Improving Library Website Content and Design: A Case Study." *Library Philosophy and Practice 2007: LPP Special Issue on Libraries and Google*. http://www.webpages.uidaho.edu/~mbolin/fang.htm.

Franklin, Brinley, Martha Kyrillidou, and Terry Plum. 2009. "From Usage to User: Library Metrics and Expectations for the Evaluation of Digital Libraries." In *Evaluation of Digital Libraries: An Insight Into Useful Applications and Methods*, 17–40. Chandos Publishing.

Galvin, T, and Allen Kent. 1977. "Use of a University Library Collection." *Library Journal* 102 (20): 2317–2320.

Giglia, E. 2010. "The Impact Factor of Open Access Journals: Data and Trends." In Helsinki. http://hdl.handle.net/10760/14666.

Gwenda, Thomas. 2007. "Evaluating the Impact of the Institutional Repository, or Positioning Innovation Between a Rock and a Hard Place." *The New Review of Information Networking* 13 (2) (November): 133–146. doi:10.1080/13614570802105992. http://www.tandfonline.com/doi/abs/10.1080/13614570802105992.

Harnad, Stevan, and Tim Brody. 2004. "Comparing the Impact of Open Access (OA) Vs. Non-OA Articles in the Same Journals." *D-Lib* 10 (6). http://www.dlib.org/dlib/june04/harnad/06harnad.html.

Harnad, Stevan, Tim Brody, Francois Vallieres, Les Carr, Steve Hitchcock, Yves Gingras, Charles Oppenheim, Chawki Hajjem, and Eberhardt Halif. 2008. "The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update." *Serials Review* 34 (1): 36–40. doi:10.1016/j.serrev.2007.12.005. http://eprints.soton.ac.uk/id/eprint/265852.

International Coalition of Library Consortia. 1998. "Guidelines for Statistical Measures of Usage of Web-based Indexed, Abstracted, and Full Text Resources." http://www.library.yale.edu/consortia/webstats.html.

Jewell, Christine, William Oldfield, and Sharon Reeves. 2006. "University of Waterloo Electronic Theses: Issues and Partnerships." *Library Hi Tech* 24 (2): 183–196.

Jin, Yi. 2004. "The Development of the China Networked Digital Library of Theses and Dissertations." *Online Information Review* 28 (5): 367–370.

Jones, Richard, Theo Andrew, and John MacColl. 2006. "Case Study: The Edinburgh Research Archive." In *The Institutional Repository*. Oxford: Chandos Publishing.

Kamberelis, George, and Greg Dimitriadis. 2005. “Focus Groups: Contingent Articulations of Pedagogy, Politics, and Inquiry.” In *The Sage Handbook of Qualitative Research*. 3rd ed. Sage Publications.

Kim, Hyun Hee, and Yong Ho Kim. 2007. “An Evaluation Model for the National Consortium of Institutional Repositories of Korean Universities.” *Proceedings of the American Society for Information Science and Technology* 43 (1-19) (October). doi:10.1002/meet.1450430176.

Kyrillidou, Martha, Terry Plum, and Bruce Thompson. 2010. “Evaluating Usage and Impact of Networked Electronic Resources Through Point-of-Use Surveys: A MINES for Library Study.” *The Serials Librarian* 59 (2): 159–183.

Lawrence, S. 2001. “Free Online Availability Substantially Increases a Paper’s Impact”, May. http://www.nature.com/nature/debates/e-access/Articles/lawrence.html.

Linn, Mott. “Cost-benefit Analysis: a Disparagement of Its Misuse and Misexplanation.” *The Bottom Line: Managing Library Finances* 22 (3): 82–85. doi:10.1108/08880450910999640.

Lippincott, Joan K., and Clifford Lynch. 2010a. “ETDs and Graduate Education: Programs and Prospects .” *Research Library Issues* 270: 6–15. http://publications.arl.org/rli270/7.

Lippincott, Joan K., and Clifford A. Lynch. 2010b. “ETDs and Graduate Education: Programs and Prospects.” *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC* 270 (June): 6–15. http://www.arl.org/resources/pubs/rli/archive/rli270.shtml.

Lomax, Joanne. 1997. “The Work of the University Theses Online Group: Report of a Survey and Seminar.” *Program: Electronic Library & Information Systems* 4: 377–382.

MacColl, John. 2002. “Electronic Theses and Dissertations: a Strategy for the UK.” *Ariadne* 32. http://www.ariadne.ac.uk/issue32/theses-dissertations.

Macduff, Colin. 2009. “An Evaluation of the Process and Initial Impact of Disseminating a Nursing E-thesis.” *Journal of Advanced Nursing* 65 (5): 1010–1018.

Mahoney, J, and G. Goertz. 2006. “A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research.” *Political Analysis* 14: 227–249. doi:doi:10.1093/pan/mpj017.

Merton, Robert K. 1952. *Focused Interview.* New York: Columbia University.

Miller, Rush, and Sherrie Schmidt. 2001. “E-Metrics: Measures for Electronic Resources.” In *the 4th Northumbria International Conference on Performance Measurement in Libraries and Information Services*. Pittsburgh, PA: ARL. http://www.arl.org/bm~doc/miller-schmidt.pdf.

Piorun, May, and Lisa A. Palmer. 2008. “Digitizing Dissertations for an Institutional Repository: A Process and Cost Analysis” 96 (3): 223–229. doi:10.3163/1536-5050.96.3.008.

Plaza, B. 2009. “Monitoring Web Traffic Source Effectiveness with Google Analytics: An Experiment with Time Series.” *Aslib Proceedings* 61 (5): 474–482. doi:10.1108/00012530910989625.

Plaza, Beatriz. “Google Analytics: INTELLIGENCE FOR INFORMATION PROFESSIONALS.” *Online* 34 (5): 33–37.

Pringle, James. 2004. “Do Open Access Journals Have Impact?” *Nature*. http://www.nature.com/nature/focus/accessdebate/19.html.

Schopfel, Joachim and Cherifa Boukacem-Zeghmouri. 2010. “Assessing the Return on Investments in GL for Institutional Repositories.” In *Grey Literature in Library and Information Studies*, 1–20. 1st ed. De Gruyter.

Shim, Wonsik, Charles R. McClure, Bruce T. Fraser, John Carlo Bertot, Arif Gagli, and Emily H. Leahy. 2001. *Measures and Statistics for Research Library Networked Services: Procedures and Issues – ARL E-METRICS PHASE II Report*. Washington, DC: Association of Research Libraries. www.arl.org/bm~doc/phasetwopreface.pdf.

Swain, Dillip K. 2010. “Global Adoption of Electronic Theses and Dissertations.” *Library Philosophy and Practice*: 418. http://digitalcommons.unl.edu/libphilprac/418.

The Royal Society. 2012. *Science as an Open Enterprise: Costs of Digital Repositories*. http://royalsociety.org/policy/projects/science-public-enterprise/digital-repositories/.

Vijayakumar, J. K., and Manju Vijayakumar. 2007. “Importance of Doctoral Theses and Its Access: A Literature Analysis.” *The Grey Journal* 3 (2): 67–75.

Wacha, Megan, and Meredith Wisner. 2011a. “Measuring Value in Open Access Repositories.” *The Serials Librarian* 61: 377–388. doi:10.1080/0361526X.2011.580423.

———. 2011b. “Measuring Value in Open Access Repositories.” *The Serials Librarian* 61 (3-4): 377–388. doi:10.1080/0361526X.2011.580423.

Web Analytics Association. 2008. *Web Analytics Definitions*. http://www.digitalanalyticsassociation.org/resource/resmgr/PDF_standards/WebAnalyticsDefinitions.pdf.

Zhang, Yin, and Kyiho Lee. 2001. “Features and Uses of a Multilingual Full-text Electronic Theses and Dissertations (ETDs) System.” In *Proceedings of National Online*, 555–566. Medford, NJ: Information Today.

Zuccala, Alesia, Charles Oppenheim, and Rajveen Dhiensa. 2008. “Managing and Evaluating Digital Repositories.” *Information Research* 13 (1).

5 Managing the Lifecycle of ETDs: Curatorial Decisions and Practices

Bill Donovan (Boston College)

Topics Covered

- Causes of short-term and long-term risks to the accessibility of ETDs.
- Guidance for policies to promote the longevity of ETDs.
- Issues specific to complex ETDs with dependencies on multiple objects.
- Strategies to ensure the stability of content during migrations.

5.1 Introduction

Inevitably, and for a variety of reasons, electronic thesis and dissertation (ETD) collections will need to be moved, updated, or otherwise modified in order to accommodate technological changes, to eradicate errors discovered belatedly (e.g., after ingest), or to enhance the collection with new content (e.g., to improve or extend metadata). Potentially, over time, a multiplicity of "versions" will be created on different storage media, with different file formats, and with deliberate or inadvertent variations in content. Among the many challenges of digital preservation will be: to decide what to preserve; how to proactively manage any changes to ETD collections; and how to adequately document these changes so that there will always be a continuous record of the entire lifecycle of the digital content. One reason to be hopeful about the tractability of these daunting challenges is the ongoing active collaboration among members of the digital preservation community to achieve sustainable systems, based on best practices and standards.

Until the late 20th century, a thesis or dissertation consisted of tangible media; in other words, a physical form such as printed text-on-paper or microfilm. While discovery or access to these works was occasionally problematic, there was little doubt that in years to come that the same physical medium would still be there "on the shelf" for future scholars to find, access, and read. In contrast, today's ETDs are much easier to find and access, often involving a fast online search followed by a download of a single digital file from an institutional repository or from a database. An interlibrary loan of an ETD might entail an email with the ETD as an attachment. However, the expectation that an ETD will still be there for future scholars to find, access, and read hinges on actively taking steps to ensure that this will be true; "benign neglect" does not work for digital content.

This guidance document considers some of the risk factors for ETDs, specifically those related to file formats, migration, and versioning. It provides advice regarding how best to understand these risks and mitigate them through implementing procedures and policies that explicitly seek to ensure the longevity of these ETD collections.

5.2 Overview of Some Risk Factors for ETDs

5.2.1 File Formats: Making Sense Out of the Bits

File formats vary in their complexity, their level of use, and their degree of openness. Each of these characteristics has implications for the future renderability of digital objects stored in any given file format. File formats may be based on an “open standard” or on a closed, proprietary standard. Open standards provide open documentation that can be used by technologists to construct a reader to render objects stored in those standards. Proprietary standards do not provide this documentation in any publicly available way; instead, only the owner of the standard knows how to render the objects stored in that format. Formats that are based in an open standard may be accessed without the purchase of software through freely available file readers built for that standard. Formats that are based in a proprietary standard often cannot be displayed without the purchase of a licensed file reader.

ETD programs should consider these factors carefully. Whenever it is possible to use non-proprietary file formats, especially those that are in widespread use, doing so will promote future access and usability in two ways. First, users will be able to render the files with freely available readers. Second, as these formats become incompatible with new versions of software (due to lack of backward compatibility support or, as some companies go out of business or abandon certain product lines, their old file formats will no longer be supported), open standards are available publicly so that technologists can continue to create tools to enable these files to render. This longevity problem is well known.[1]

5.2.2 Disk Failures, Missing Files, Corrupted Files, and Other Calamities

Inevitably storage media develop problems over time. For example an entire hard drive can crash with a total loss of files or perhaps one or more files become corrupted (aka “bit rot”) so that they are no longer readable. Precautions against such calamities often consist of backing up the files, preferably on a storage device in a different location. But in many cases, backups are insufficient because they merely replicate the problem (e.g., a file degrades without anyone noticing, and the degraded file is replicated in the backups) or are stored too closely to the original (e.g., the backup copies are within the same geographic area and/or power grid, and thus are prone to the same physical risks as the original, e.g., flooding, physical attack, or prolonged power failure).

A more proactive process is needed to detect when files are corrupted or lost, while there is still time to repair or replace files that have been compromised. The detection part can be accomplished with a “fixity-checking” tool that computes a unique digest for each file that depends on the exact bit stream contained in the file. If even a single bit changes, the recomputed digest (“checksum”) will differ,

[1] See http://www.dpconline.org/events/previous-events/306-digital-longevity.

signaling a loss of file integrity. Given at least one intact back-up copy, the corrupted file can be then be replaced.[2]

In distributed digital preservation, multiple copies of the file are stored over a wide geographic area, and the checksums associated with these copies are regularly checked, e.g. using LOCKSS,[3] so that repairs to corrupted files can be made when necessary. Storing the copies in different geographical zones (ideally, at least 200 miles apart, in different weather zones, power grids, etc.), also safeguards against loss due to localized physical calamities such as hurricanes, power failures, and fires.

5.2.3 Technological Obsolescence

Even preserving the bit streams of files and moving them periodically to newer storage media does not guarantee future meaningful access to the digital content. As hardware and software evolve, these changes in the computing environment will impact the viability of our ETD collections. To accommodate these changes, one might consider either (or both) of two strategies: emulation and/or migration, as discussed below.

It may be possible in the future to emulate today's computing environment, in essence reconstructing the user experience of the past. For example, embedding fonts in a PDF file is one step undertaken to ensure that the text in an ETD will still look the same to future generations of users. Another step is to be able to run the obsolete file reader application that was able to render those fonts, but on a future computer whatever its operating system. This might require emulating obsolete hardware (such as an older CPU bus architecture) with software running on more recent hardware, so that the older file reader application can render the original version of the archival file.

Besides or in addition to emulation, it may also be prudent to migrate archival files to next generation file formats and to add metadata to record these "lifecycle events," and take any other steps required to ensure that ETDs can still be discovered, accessed and read with future "computers." Examples might be converting BMP files to TIFF or RTF files to PDF. The purpose of migration is to always have a version that can be handled by the current computing environment. The topic of file format migration is discussed more extensively later in this document.

5.2.4 Loss of Relationships among Digital Content

Some digital content consists of multiple components that collectively comprise an intellectual entity, e.g. a website or the output of a research study. If any of these components are missing or the relationships among these components are no longer retained, the semantic integrity of the content is lost. For these "complex content objects," structural and other metadata (as well as the naming of directories and files) are often used to convey the interrelatedness of the components and to keep the components and the metadata together as a single group. As in *Figure 5-1* below, for a chemistry dissertation there may be laboratory notebooks and measured spectra whose relationship to the main

[2] A detailed discussion of fixity-checking tools and auditing approaches can be found at: http://www.metaarchive.org/portal/integrity-checking.

[3] See http://www.lockss.org/about/what-is-lockss/.

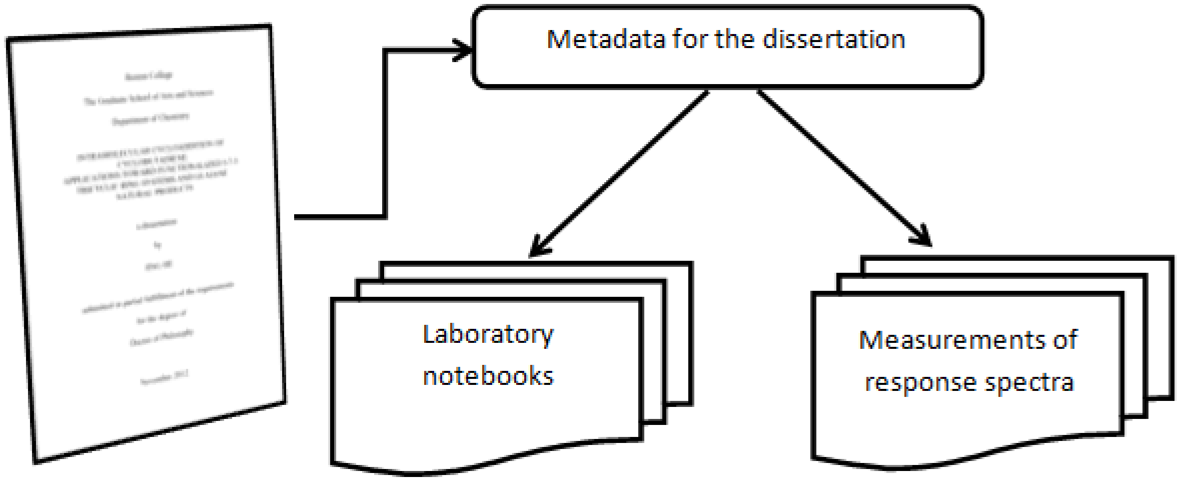

Metadata for the dissertation
Laboratory notebooks
Measurements of response spectra

Figure 5-1. The metadata accompanying a dissertation may include references to supplemenetary files such as digitized laboratory noteb and sprectral response measurement data.

text could be explained in the metadata so that a particular section of the dissertation describing an experiment can be linked with the pertinent sections of the laboratory notebooks and the associated recordings of spectra. Links to resulting publications, or at least their citations, might also be added subsequently to the metadata.

While ETDs are still predominantly text-based objects, gradually more multimedia are being used, and ETDs as digital objects are otherwise becoming more complex. For example, it is not that unusual for an ETD submission to include a PDF representing the main body of text, an xml file with the various types of metadata, and supplementary files that contain related information that is not part of the dissertation itself (spreadsheets, maps, audio and video files, copyright permissions etc.). Some supplementary files may be invoked or otherwise be referred to from within the dissertation. All of these files together constitute an ETD; the relationships among these files are important to the meaning of the scholarly work. Retaining these relationships is part of the challenge of digital preservation. As files are stored and as they are migrated to newer file formats, these relationships must be kept intact.

The following sections of this document will now explore all of the above issues in greater depth.

5.3 Data Wrangling: Organizing Digital Content Prior to Ingest

The future usability of any preserved digital content will depend in part on how well organized the body of content was when it was first ingested. So, it is vitally important at the outset to establish and follow a logical set of principles and conventions that inform the organization of the content and will be readily understood in the future. In some cases, this may require remediation of flawed legacy practices (Halbert et al. 2008).

The guidelines and examples that follow are not prescriptive but are meant to exemplify a thought process that might lead to an optimum set of practices that should then be codified in the policies and the procedures that undergird any mature ETD program.

5.3.1 File/Folder Naming Principles and Conventions:

File/folder names should be unique and follow documented conventions to ensure consistency and ease of use. File names do not take the place of metadata and should be simple and straightforward.

Possible conventions:[4]

- Use lowercase letters of the English alphabet and the numerals 0 through 9.
- Avoid punctuation marks other than underscores or hyphens.
- Do not use spaces.
- Use leading zeroes (where needed) to maintain equally-long numerical strings in order to facilitate sorting.
- If dates are included, adhere to the ISO 8601 convention, YYYYMMDD.
- Keep file/folder names brief. Long names may not be as portable.

Examples for files:

- solere-chemins_0091.tif
- nl-mix-xul_20120928.pdf
- brooker-0101b.jp2

Examples for folders (indentations being used to denote folder hierarchy):

```
2005……………………………. year

|---q2…………………………. second quarter

    |---embargo………… no access to full text

    |---local………………. campus-wide access only

    |---public…………….. open access

    …
```

5.3.2 Normalization

While many institutions mandate what format is to be used for an ETD, there may remain some degree of ambiguity. For example, even when a PDF file format is required, both the PDF and PDF/A formats may be considered acceptable; there may be no explicit requirement as to which version of PDF is used.

[4] Adapted from Appendix I of Bogus et al. (2013): www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations.

For any files that are included as supplementary files to the ETD itself, even fewer restrictions may apply. For these files, merely requiring that non-proprietary file formats be used may be insufficient;given the wide variety of non-proprietary formats, one or more media-specific sustainable (amenable to preservation) formats (see *Table 5-1*) should be recommended. For legacy files, it may be worthwhile to "normalize," i.e. to convert files from their original proprietary formats to more sustainable formats, which are preferred for archival purposes.[5]

Content	Software	Format	Sustainable format	Notes
text	Microsoft Word	.doc, .docx	.pdf	Preferably, PDF/A or PDF version 1.7. See *Appendix A*. For preservation, all fonts should be embedded in the PDF and all security features should be disabled.
	LaTeX	.tex		
image	Adobe Photoshop	.psd	.tif	
			.jp2	JPEG2000[6]
			.jpg	
tabular	Microsoft Excel	.xls, .xlsx	.ods	Spreadsheet for the ODF (Open Document Format)[7]
			.csv	Comma (Character) separated values
			.tsv	Tab separated values
audio	various	.mp3	.wav (BWF)	Broadcast Wave[8]
		.aif		
		.m4a		
video	various	various	.dv	Video is an area of ongoing development and debate in digital preservation.[9]
			.avi	
			.mov	

Table 5-1. A sample list of content types and format normalization choices.[10]

Table 5-1 is not meant to be prescriptive, comprehensive, or definitive – especially given the continuing evolution of file formats, codecs, etc. Instead, it illustrates how one might think about the challenge of promoting file formats that will make it easier to accomplish not only the bit preservation of original

[5] Regarding the sustainability of formats please see: http://www.digitalpreservation.gov/formats/index.shtml. Regarding what file formats are currently in use, consult PRONOM http://www.nationalarchives.gov.uk/PRONOM/; and UDFR http://www.udfr.org/.

[6] See Buckley, 2008 and Lowe and Bennett, 2009.

[7] OpenDocument Format (ODF): http://www.opendocumentformat.org/.

[8] See http://www.iasa-web.org/tc04/key-digital-principles.

[9] See Jones 2011 http://blogs.loc.gov/digitalpreservation/2011/07/whither-digital-video-preservation/ ; and Murray 2013 http://blogs.loc.gov/digitalpreservation/2013/10/one-format-does-not-fit-all-fadgi-audio-visual-working-groups-diverse-approaches-to-format-guidance/.

[10] Please note that a similar but more extensive table (Gregory, 2011), including a listing of tools (open source and otherwise) that may be used to effect these file format migrations, can be found at: http://www2.archivists.org/sites/all/files/LGFinal_0.pdf.

files but also the migration to newer file formats that inevitably supercede file formats that become obsolete (are no longer supported by software applications and/or operating systems). The proposed sustainable formats have been chosen with preservation in mind, as opposed to the derivative files that are generated for "delivery" and "use." The rationales for the choices in this table and their associated caveats are discussed below.

NOTE: All files should be checked for viruses and their formats validated[11] before ingest.

5.3.2.1 Rationale/caveats for text

As discussed in Appendix A, there are several variants of the PDF file format and of course multiple versions as well. Version 1.7 of the PDF file format is an ISO standard[12], thereby making it in effect an open source (non-proprietary) file format that recommends its use for archival purposes. PDF/A is also an ISO standard that was originally developed as an "archival" version of PDF. Unlike "full PDF," PDF/A does not support multimedia or hyperlinking. There is currently no consensus as to which of these two versions of PDF is better for digital preservation; both are being used at different universities. Meanwhile newer versions of PDF and PDF/A have been under development and may yield even better alternatives.[13] In addition to saving the PDF file, it may also be worthwhile to save a precursor file such as LaTeX, in case the software for converting to PDF (or some other file format) is improved in the future.

5.3.2.2 Rationale/caveats for images

TIFF is widely used for archiving digital still images, and its internal tags are employed to store key technical metadata.[14] In saving images to TIFF format, one can choose either no compression or lossless compression (e.g., LZW) in order to preserve images at their highest quality. Arguably[15], JPEG2000 is an attractive alternative to TIFF when preserving large images for which high-resolution detail is important to retain. JPEG2000 can serve as a delivery format as well as an archival format; when rendered with suitably-designed viewing software, JPEG2000 makes it easier (than TIFF) to navigate across a large image (such as a map, newspaper, or an elephant folio book) and then zoom in and out to perceive the fine details. Lossless or no compression is supported as part of the JPEG2000 format, but a modest amount of compression that might in effect be "visually lossless" is an option to consider (Buckley 2008). Finally, for images that do not need to be preserved at an archival quality level, simple JPEG may be sufficient and even preferable since it will cause less of an increase in file size.

[11] Regarding testing their validity please see: JHOVE2, https://bytebucket.org/jhove2/main/wiki/documents/Abrams_a70_pdf.pdf; DROID https://github.com/digital-preservation/droid#readme; and FITS http://code.google.com/p/fits/.

[12] See http://www.digitalpreservation.gov/formats/fdd/fdd000277.shtml.

[13] Information about the recently published standard PDF/A-3 can be found here: http://www.digitalpreservation.gov/formats/fdd/fdd000360.shtml.

[14] See ANSI/ISO Z39-87 2006, http://www.niso.org/apps/group_public/download.php/6502/Data Dictionary - Technical Metadata for Digital Still Images.pdf.

[15] As witnessed by a recent (March 2013) thread by members of the Digital Curation Google Group: https://groups.google.com/forum/?fromgroups=#!topic/digital-curation/jX1MELYvpKE.

5.3.2.3 *Rationale/caveats for tabular*

OpenDocument (OD), based on an ISO standard, includes an xml-based file format for spreadsheets (with a file extension of ".ods"). Given the interoperability that now exists with commercial (proprietary) spreadsheet application software, one may be able to convert from Excel format (for example) to OD format to produce an archival version of a spreadsheet. However, be aware that differences between file formats may result in some losses in information or functionality.[16] Alternatively, one could convert a spreadsheet to PDF format; again, losses or objectionable changes in layout might occur.

5.3.2.4 *Rationale/caveats for audio*

The WAV (.wav) format is an established preservation standard for digitization of analog audio and migration of born-digital audio, specifically as Broadcast WAV (IASA). Even though WAV is a proprietary format (developed by Microsoft and IBM) it has been adopted as the de facto standard rather than any open formats (such as FLAC[17]) presumably because of its widespread use and its portability.[18]

5.3.2.5 *Rationale/caveats for video*

There is as yet no consensus regarding what file format to use for preserving video. One candidate for a future standard is lossless JPEG2000 wrapped in MXF.[19] Video is problematic in that many video formats are already compressed. That said, the best practice at this point seems to be to keep a digital video format in its native wrapper (e.g. .dv, .avi, .mov) and to do no further compression (as in the case of reformatting mini-DV to .dv files without transcoding). Even compressed video files can be very large. For example, a mini-DV to .dv transfer could yield roughly 13 GB per hour of video. Whether or not the resulting file is prohibitively large depends on how much space is available. But even this is seen as an "interim" approach (as is the use of other proprietary video wrappers such as .mov and .avi) until the move to a standard such as lossless JPEG2000/MXF (as employed by the Library of Congress) becomes more widely adopted and more within reach for smaller institutions. This state of affairs and ongoing efforts by FADGI has been summarized by Kate Murray (2013).

5.3.3 Pre-ingest checklist

Finally, it may prove helpful to devise your own checklist of the suite of actions that you wish to undertake before each ingest. The checklist might contain the following tests:

- File name adheres to local conventions.
- File is free of any virus.
- File format is valid.
- File format has been normalized, if necessary.
- Checksum has been computed.
- Metadata has been vetted.

[16] Please see: http://office.microsoft.com/en-gb/excel-help/differences-between-the-opendocument-spreadsheet-ods-format-and-the-excel-xlsx-format-HA010355787.aspx?CTT=5&origin=HA101878944.

[17] FLAC: http://flac.sourceforge.net/.

[18] With appreciation to my colleague John Kearney for the information that he provided regarding audio and video.

[19] See http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml.

Tools for some of these actions already exist and some are discussed elsewhere in this document. Once you decide on your preferred tool and workflow for each action, you should document your choices. When choosing tools, be on the lookout for open source tools, especially those that will help to automate workflows. One example is Fixity,[20] a tool for scheduling checksum computations and automated comparisons that sends reports after each check to a curator's email inbox. As with any such tool, you will need to carefully assess whether it serves your purposes.

5.4 Complex Content Objects

An "ETD" consists of more than just a single file that contains the "full text" of the thesis/dissertation itself. Each ETD is a complex content object, or "intellectual entity" in PREMIS terminology,[21] comprised of several components. Some or all of these components could be encoded as separate files. Or, some could be embedded within the full text file. For example, a link can be embedded within a PDF which when clicked opens an audio file with an appropriate media player. At the very least, there would be a metadata component (typically a separate xml file) as well as the full text file. There might also be components consisting of data, audio, video, maps or other graphics. Each may be a separate file. And, there could be permissions for using copyrighted material or for granting permission for Open Access and/or for digital preservation (including future migration to newer file formats). In effect, a single package of these interrelated files is needed to keep them all together (virtually perhaps) as one complex content object.

Embedding multimedia components within the full text might seem advantageous in that they would then be inseparable. However, when the time comes that it is necessary to migrate either the full text itself or one of the multimedia components, having separate files would greatly simplify matters. For that reason, it is better for an ETD author to embed links to multimedia within the full text and include the multimedia as separate files. If instead, the author had embedded the multimedia within the full text, then an ETD curator should consider whether or not to remediate the file, i.e. by replacing or supplementing the embedded multimedia with links to multimedia as separate files. Ultimately, the "packaging" of an ETD and all of its constituent parts (metadata, fonts, data, multimedia, etc.) must be accomplished in such a way that access to the ETD in the future will encompass all of its intellectual content and functionality, as follows.

5.4.1 Metadata

An ETD-specific item-level metadata schema ETD-MS has been developed by members of the Networked Digital Library of Theses and Dissertations (NDLTD).[22] ETD-MS, based on Dublin core, is primarily a descriptive schema; it does not include elements for lifecycle management events or structural relationships. Instead, elements pertaining to lifecycle management can be found in PREMIS; they are needed to record the status of content at the time of ingest, e.g., fixity checksums, as well as any subsequent actions undertaken on behalf of preservation, e.g. format migration. The ETD-MS

[20] Fixity: http://www.avpreserve.com/avpsresources/tools/.

[21] See http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/premis-data-dictionary#2.

[22] Example of ETD-MS metadata: http://dcollections.bc.edu/webclient/MetadataManager?pid=139660&descriptive_only=true.

dc.rights element can be used to record any Creative Commons licensing options. Additional metadata elements as in a METS document are needed to record relationships among groups of files that constitute a complex content object. In effect, metadata can provide the "glue" that binds these files together. A more detailed discussion of metadata issues may be found in the guidance document *Metadata for ETD Lifecycle Management*.

5.4.2 Fonts

Whether fonts are selected arbitrarily or very deliberately by an ETD author, they are part of their ETD and as such should be preserved. Currently, the embedding of all fonts can be easily accomplished when converting, for example, from Microsoft Word to Adobe Acrobat PDF. In fact, having all fonts embedded is sometimes an explicit requirement (e.g., of UMI ProQuest). However, many ETD authors succeed in embedding only some fonts, but not all. Recently, it has become possible to "fixup" these ETDs with Adobe Acrobat Pro's "preflight fixups embed fonts" function, which embeds any missing fonts, as long as those fonts are available among the local computer's system fonts. Ideally, fonts should be embedded while using the same operating system (better yet, the same computer) as was used when authoring the ETD in order to ensure that the exact same fonts (or subsets thereof) are embedded within the PDF.[23]

5.4.3 Hyperlinks

More and more frequently, hyperlinks are included within ETDs and they will remain active within some versions of PDF (depending on the conversion-to-PDF settings). Even if these links eventually "break" (aka "link rot"), arguably their inclusion provides some potentially valuable information (Taylor 2012), especially if tools become available for repairing broken links. Going forward, technologies are being developed to avert and to mitigate link rot, e.g. WebCite and PermaCC.[24] Part of the solution will be to employ persistent identifiers, discussed next.

5.4.4 Persistent Identifiers

For the ETDs themselves, any of several schemes and services may be employed to ensure that they can be located in the future despite their relocation to different file servers. These are listed in *Table 5-2*. The process of minting and maintaining persistent identifiers can be automated with tools such as EZID.[25]

Schema	Example	Documentation
Handle	http://hdl.handle.net/2345/1328	http://www.handle.net/introduction.html
ARK	http://digital.library.unt.edu/ark:/67531/metadc115051/	https://confluence.ucop.edu/display/Curation/ARK
DOI	doi:10.1016/S0065-2830(06)30001-3	http://www.doi.org/
PURL	http://purl.oclc.org/OCLC/RSPD	http://purl.oclc.org

[23] A more detailed description of this fixup procedure can be found at: http://blogs.adobe.com/acrobatforlifesciences/2008/09/reembedding_fonts_in_a_pdf/.

[24] See http://www.webcitation.org/ and http://www.perma.cc/.

[25] See http://n2t.net/ezid.

Table 5-2. Persistent URL Schema

5.4.5 Multimedia

The use of multimedia (audio, video clips, etc.) in ETDs has been gradually increasing, as is evident in many of the recent ETDs recognized by the NDLTD for their exemplary use of the ETD format.[26]

As explained above, multimedia components should consist of supplementary files that are separate from the main ETD text, thereby making it possible to migrate these multimedia components to newer file formats independently of the full text PDF. However, if this migration requires a new extension for the filename, then the embedded link will also need to be modified so that it points to the new multimedia file. Both files will then need to be handled together as a new version of the ETD. This type of joint migration of a complex content object is a prime candidate for automation and, hopefully, appropriate tools will be forthcoming if they are not already available. Fortunately, open source media players are becoming available that will be capable of playing non-proprietary multimedia file formats thereby ensuring the future usability of the content.[27]

5.4.6 Research Data

For some ETDs, the research data upon which the ETD was based (e.g., from surveys or from measurements) might also be considered preservation-worthy. Accordingly, one might preserve such data as part of an "ETD package" of files that are stored in a preservation network, or one might archive the data in a separate data repository that can be referenced within the ETD package. As in the example depicted earlier, for a dissertation in chemistry, laboratory notebooks and measured spectra could be converted to archival files that could go into an open access data repository; links to the data repository from the ETD metadata would then enable researchers to access these research data in the future. Data preservation is currently an area of active investigation. For example, Buneman et al. (2004) have described some of the challenges when archiving scientific data. Lemire and Vellino (2011) have proposed a model "for managing and mechanizing the curation of research data." Conway et al. (2011) have emphasized the value of preserving research data in such a manner that it may be possible to re-analyze the data in the future.

5.4.7 All-in-One

Rather than having dispersed files that could potentially get separated in the future, having an ETD and all of its associated information be in one self-contained bundle might prove to be more amenable to preservation. This is the premise of an approach being tested at Virginia Tech (Park et al. 2010), using HTML5, which "allows a single file to encode multiple media types and support linking among those."

Another possible approach is to utilize a technology developed by the Library of Congress and the California Digital Library called BagIt.[28] This file-packaging specification was designed for transferring or storing files, either between disks or over the network. It employs a hierarchical file structure and has a

[26] NDLTD. "NDLTD ETD Awards - 2013 Winners Announced," http://www.ndltd.org/events/news/ndltdetdawards-2013winnersannounced.

[27] For example, see: http://www.videolan.org/vlc/index.html.

[28] http://www.digitalpreservation.gov/news/2008/20080602news_article_bagit.html.

built-in inventorying system that creates a manifest of whatever is put into the bag and uses that manifest to ensure that the same files can be removed intact from the bag at some later time, i.e. with no changes to the file bit streams. So, one could conceivably put all files related to a single ETD into a single bag to make them inseparable and to have a straightforward check on file integrity, by comparing checksums (original versus current).

5.5 Migration Scenarios

5.5.1 Media Migration

As media storage technologies evolve, it will be prudent to transfer digital collections to newer storage media. To ensure that all of the digital content is transferred without error, fixity checks will need to be done immediately before and after the transfer in order to compare the original files with the newly-transferred files. Absent any discrepancies, it would then be safe to decommission the original storage media. A workflow for automating such fixity checks would include tools that both compute checksums or other integrity metrics as a batch process and then also compare the newly-computed metrics against their expected values and report out any discrepancies.[29] For example, the open source tool Fixity available from AVPreserve lets you do just that and emails you a report indicating what if any discrepancies were discovered. A more limited but potentially useful tool is FileAnalyzer which batch-computes checksums but does not then perform comparisons.[30]

Or, finally, one could use Unix commands such as *md5sum* and *diff* to compute and compare checksums respectively, provided that the formatting of the checksum files is consistent.

5.5.2 Format Migration

As file formats evolve and older formats approach the point of obsolescence, it becomes critical to migrate the content of the files to those formats that are most logically the next generation equivalent file, while it still possible to do so. Implicit in this recommendation are several assumptions: that you have some reason to believe that a specific file format is about to become obsolete; that you are able to predict the format that will replace it; and that tools exist with which to perform the migration and efficiently enough to handle large sets of files.[31]

The issue of tools has been discussed in detail by Gregory (2010); her emphasis is on the use of Open Source tools to perform file format migrations, but recognizes that in some cases proprietary software is needed to effect a particular migration and that in some cases changes in layout, or losses in metadata,

[29] If you use an ftp application to transfer files from a DOS-based system to a Unix-based system, be aware that you should transfer xml files (for example) as binary files; otherwise the checksums may not agree (because the “carriage-return/line-feed” in DOS can change into just “line-feed” in Unix).

[30] See http://blogs.archives.gov/online-public-access/?p=6270. Note: this tool overwrites the checksum if it encounters another file with the same filename in a different directory.

[31] Extensive information about the sustainability of individual file formats can be found at: http://www.digitalpreservation.gov/formats/index.shtml. Additional resources for assessing the viability of specific file formats (e.g. a list of what applications can still read and write such files) can be found at: (PRONOM) http://www.nationalarchives.gov.uk/PRONOM; and (UDFR) http://www.udfr.org/.

or losses in functionality will occur given the tools that currently exist. Many concrete examples of attempted file format migrations are given along with insightful assessments of each outcome.

5.5.3 Fixity-Checking across Formats

As part of any file format migration, fixity checking will need to be performed in order to establish the checksums associated with the new "generation" of files. Assuming one starts with multiple copies (replications stored at distributed locations) of files with checksums that have been checked for agreement with one another, one should end up with multiple copies of the newly-migrated files with checksums that also agree with each other but not, of course, with those of the original files. Moreover, if multiple copies of each file are migrated independently, and the new checksums all agree, that would indicate success.

5.5.4 When to Migrate to a Newer File Format

While widely-adopted non-proprietary file formats such as TIFF and PDF are not likely to become obsolete any time soon, one might anticipate that almost any file format may one day be superceded by another format with better functionality or smaller file sizes or some other technical advantage. So far, however, the file formats typically employed with ETDs have not been in danger of obsolescence. Nevertheless, the eventual challenge will be to predict when a file format is in danger of becoming obsolete far enough in advance to plan for migration to a newer more stable file format. This predictability may depend on having a technology watch system in place, or subscribing to a technology watch service for example the DPC Technology Watch Series (Todd, 2009).

Two possible strategies for deciding when to migrate files to a newer file format are "preemptive" format conversion versus "on-demand" format conversion. With the former strategy, one would predict what the next archival format is going to be and then batch-convert all soon-to-be-obsolete files into that format. With the latter strategy, one would wait until the file whose format is outdated is going to be accessed, and then convert only that specific file. At this point, having not yet needed to migrate any files, it is not clear which strategy will ultimately prove to be most efficient and sustainable. However, if the "window" (time period) during which migration is feasible is brief (*Preservation Management of Digital Materials: The Handbook*, 2009), then the on-demand strategy may prove to be untenable.

5.5.5 Versioning

Versioning is the process of storing multiple versions of a file in order to save its change history. This enables a content producer to know that changes made to a preserved file, intentional or otherwise, will be saved in parallel within a preservation system such that any/all versions of that file may be retrieved by the producer in the event of a content restoration session. However, one will need to consider the cost implications of saving all intermediate versions; given the potential for escalating storage costs (Rosenthal et al. 2012), a purging /de-accessioning mechanism may be essential.

Note that versions can be the result of generating new files to accommodate internal content changes but without changing file format. And, versions can be the result of file format migration. Regarding the former, one possible strategy is to maintain a "revision history" as part of the metadata, but not save all of the intermediate files. Regarding the latter, one possible strategy would be to save both the original archival files in order to keep open the option known as emulation (discussed above) and only the most recently migrated archival file format, but not any of the intermediate versions that might have been generated along the way.

5.5.6 Migration from One Repository to Another

As with file formats and storage technologies, repository technologies may also become obsolete, warranting migration from an older to a newer or better repository platform. However, there is currently enough variation in architecture among different repositories to raise the concern that moving content from one to another may entail some loss of information. Some efforts such as Towards Interoperable Preservation Repositories (TIPR) and Repository Exchange Package (RXP) have begun to explore this issue. And, an NEH-funded research effort is currently underway to develop lightweight tools and/or methodologies for transferring (digital newspaper) content among various repositories that have different architectures.[32] Possibly, this transfer of content could be facilitated by the BagIt file packaging specification discussed earlier in this document.

5.6 Summary

This guidance document has focused on the curatorial decisions and emerging best practices upon which future access to ETDs depends. At every stage of the lifecycle of an ETD, one or more agents such as the author, ETD administrator, or digital preservation manager performs some action that may affect long-term access to an ETD. In the not so distant future, those actions will have determined whether a researcher can still find a particular ETD, whether s/he can still open the file(s), whether their viewing experience will still be faithful to the author's original work, will still encompass the totality of the scholarly work, i.e. the ETD and any supplementary material (e.g., data, maps, audio, and video) – some of which may have been added after the full text of the ETD itself was first published.

The intent of this document has been both to raise central issues such as adopting a preservation strategy (emulation and/or migration) and to delve into more granular issues such as migration across storage media, across file formats, and across repositories. Without being prescriptive, we have pointed out the need to choose file formats prudently, favoring ones that are non-proprietary and widely-adopted. At the same time, we have cautioned that in some cases, e.g. video, it is not yet clear what file format to prefer for archival purposes, and have pointed the reader to resources that may explain why picking the best file format is currently problematic. Regarding other issues such as the naming of files and folders (prior to ingest), we have offered advice confidently given the more mature documented best practices that exist.

Inevitably, this overview is incomplete, deliberately leaving out some more esoteric issues and inadvertently leaving out other potentially useful details. However, as a stake in the ground we hope this document inspires fellow members of our ETD community to comment on and debate some of these issues, perhaps thereby accelerating the emergence of best practices and widening and deepening our understanding of what it takes to curate an ETD.

[32] See http://www.metaarchive.org/projects/neh.

Finally, new standards, new tools, and new methods for achieving sustainable preservation emerge in many different venues making it important to cast a wide net for hearing about new developments. For example, the listserv ETD-L is an invaluable forum for airing issues with ETD colleagues of all experience levels, many of whom are not necessarily in agreement about the best way to run an institution's ETD program. Many other resources are available at the websites of both NDLTD and USetdA. And, finally, through the auspices of funding agencies such as IMLS, projects like this one are making it possible to compile and disseminate briefings and guidelines that explore every phase of the lifecycle of electronic theses and dissertations.

Appendix A: Further Details about Various PDF File Formats

The Portable Document Format (PDF) file format was invented by Adobe Systems Incorporated in the 1990s. As PDF became the de facto standard for documents on the web, there have been nine versions of PDF specifications. In 2006, the ISO standard ISO 32000-1:2008 (PDF version 1.7) was released as a published ISO standard. It includes all of the functionality in versions 1.0-1.6 of the PDF file format.

PDF/A

PDF for Archive (PDF/A) is a simplified version of the "full" PDF format, with fewer requirements and fewer features. It currently consists of three versions.

PDF/A-1 is based on PDF 1.4, classified as an ISO standard (ISO 19005-1:2005) for long-term preservation of electronic documents.

PDF/A-2 is based on PDF 1.7, classified as an ISO standard (ISO 19005-2:2011). The key idea of PDF/A is that it should be a 100% self-contained electronic document. All the information required for displaying the document is embedded in the file, which includes all the content (text and images), fonts and color information. Some actions and Javascript allowed in other PDF versions are not allowed in PDF/A.

PDF/A-3 is also based on PDF 1.7 (ISO 19005-3:2012). This version of PDF/A allows one to embed files of any arbitrary format inside a PDF file. However, "… the specification makes the PDF/A-3 document a "dumb" container that prohibits "actionable" access to the embedded files," (Lazorchak 2012).: http://blogs.loc.gov/digitalpreservation/2012/11/all-in-embedded-files-in-pdfa/ .

Because a PDF/A document must embed all fonts and other information for displaying the document, its file size will be larger than PDF documents without such embedded information.

Use

PDF/A has been adopted as the standard for long-term government archives in many countries, including USA federal courts, Switzerland, Austria, Germany, and the European Commission.

References

Library of Congress. PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4, http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml.

AIIM. Frequently Asked Questions (FAQs) ISO 19005-1:2005 PDF/A-1: July 10, 2006. http://www.aiim.org/documents/standards/PDF-A/19005-1_FAQ.pdf.

PDF/A – the standard for long-term archiving, http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf.

PDF/A – A Look at the Technical Side, http://www.pdfa.org/2011/08/pdfa-%E2%80%93-a-look-at-the-technical-side/.

PDF/A-3, PDF for Long-term Preservation, Use of ISO 32000-1, With Embedded Files, http://www.digitalpreservation.gov/formats/fdd/fdd000360.shtml.

Bibliography

Abrams, S., Morrissey, S., and Cramer, T., 2008. "'What? So What?' The Next-Generation JHOVE2 Architecture for Format-Aware Characterization." *iPRES Conference 2008.* https://bytebucket.org/jhove2/main/wiki/documents/Abrams_a70_pdf.pdf.

Adobe Systems. "Reembedding Fonts in a PDF." http://blogs.adobe.com/acrobatforlifesciences/2008/09/reembedding_fonts_in_a_pdf/.

ANSI/ISO. *Z39-87_2006 Data Dictionary --- Technical Metadata for Digital Still Images.* http://www.niso.org/apps/group_public/download.php/6502/Data%20Dictionary%20-%20Technical%20Metadata%20for%20Digital%20Still%20Images.pdf (last accessed 11-16-2012).

Arms, C., Fleischhauer, C., and Jones, J. 2011. "Sustainability of Digital Formats Planning for Library of Congress Collections." *NDIIPP* http://www.digitalpreservation.gov/formats/ (last accessed 11-16-2012).

Bogus, I., Blood, G., Dale, R., Leech, R., and Mathews, D. 2013. "Minimum Digitization Capture Recommendations." ALCTS Preservation and Reformatting Section: http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations (last accessed 11-06-2013).

Buckley, R. "JPEG 2000 – a Practical Digital Preservation Standard?" *DPC Technology Watch Series.* http://www.dpconline.org/docs/reports/dpctw08-01.pdf (last accessed 11-16-2012).

Buneman, P., Khanna, S., Tajima, K. and Tan, W. 2008. 2004. "Archiving Scientific Data." *ACM Transactions on Database Systems*. 20:1-39. http://homepages.inf.ed.ac.uk/opb/papers/TODS2004.pdf (last accessed 01-25-2013).

California Digital Library. "Archival Resource Key." https://confluence.ucop.edu/display/Curation/ARK.

California Digital Library. *EZID.* http://n2t.net/ezid.

California Digital Library. *Unified Digital Format Registry (UDFR).* http://www.udfr.org/ (last accessed 11-16-2012).

Caplan, P., Kehoe, W., and Pawletko, J. 2010. "Towards Interoperable Preservation Repositories: TIPR," *International Journal of Digital Curation*, pp. 34-45. doi:10.2218/ijdc.v5i1.142 http://www.ijdc.net/index.php/ijdc/article/view/145/ (last accessed 01-25-2013).

CNRI. *HANDLE.NET Services: Global Handle Registry.* http://www.handle.net/introduction.html (last accessed 11-16-2012).

Conway, E., Giaretta, D., Lambert, S., and Matthews, B. 2011. "Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context." *International Journal of Digital Curation*. 6(2):38-52. doi:10.2218/ijdc.v6i2.204. http://www.ijdc.net/index.php/ijdc/article/view/182 (last accessed 01-25-2013).

DPC. 2009. *Preservation Management of Digital Materials: The Handbook*. http://www.dpconline.org/pages/handbook/index.html (last accessed 11-16-2012).

Florida Center for Library Automation. “Towards Interoperable Preservation Repositories (TIPR).” http://wiki.fcla.edu:8000/TIPR/1 (last accessed 01-17-2014).

Florida Center for Library Automation. “Repository eXchange Package (RXP).” http://wiki.fcla.edu:8000/TIPR/21 (last accessed 11-16-2012).

Free Lossless Audio Codec (FLAC). http://flac.sourceforge.net/ (last accessed 11-16-2012).

Gregory, L. 2010. “The "M" Word": Exploring File Format Migration with Open Source Tools.” *Society of American Archivists – 2010 Research Forum*. http://www2.archivists.org/sites/all/files/LGFinal_0.pdf (last accessed 01-22-2013).

Halbert, M., Skinner, K. and McMillan, G. 2008. *“*Avoiding the Calf-Path: Digital Preservation Readiness for Growing Collections and Distributed Preservation Networks.” In: IS&T *Archiving 2008 Final Program and Proceedings*, pp.86-91.

Harvard University Libraries. *File Information Tool Set (FITS)*. http://projects.iq.harvard.edu/fits.

Higgins, S. 2007. “PREMIS Data Dictionary.” *DPC* http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/premis-data-dictionary#2 (last accessed 11-16-2012).

International Association of Sound and Audiovisual Archives. “Key Digital Principles.” http://www.iasa-web.org/tc04/key-digital-principles (last accessed 11-16-2012).

International DOI Foundation. “Digital Object Identifier.” http://www.doi.org/.

Jackson, J. “Digital Longevity: the lifespan of digital files.” *DPC.* http://www.dpconline.org/events/previous-events/306-digital-longevity.

Jones, J. 2011. “Whither Digital Video Preservation?” *The Signal*. http://blogs.loc.gov/digitalpreservation/2011/07/whither-digital-video-preservation/ (last accessed 11-16-2012).

Lazorchak, B. 2012 “All In! Embedded Files in PDF/A” *The Signal*. http://blogs.loc.gov/digitalpreservation/2012/11/all-in-embedded-files-in-pdfa/.

Lemire, D. and Vellino, A. 2011. “Extracting, Transforming and Archiving Scientific Data.” *Fourth Workshop on Very Large Digital Libraries.* arXiv:1108.4041v2 http://arxiv.org/abs/1108.4041 (last accessed 01-25-2013).

Library of Congress. 2008. “MXF File, OP1a, Lossless JPEG 2000 in Generic Container.” *Sustainability of Digital Formats*. http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml (last accessed 11-16-2012).

Library of Congress. 2009. “Library Develops Specification for Transferring Digital Content.” http://www.digitalpreservation.gov/news/2008/20080602news_article_bagit.html (last accessed 01-25-2013).

Library of Congress. 2010. “MXF PDF, Version 1.7 (ISO 32000-1:2008).” *Sustainability of Digital Formats*. http://www.digitalpreservation.gov/formats/fdd/fdd000277.shtml (last accessed 11-16-2012).

Library of Congress. *Sustainability of Digital Formats.* http://www.digitalpreservation.gov/formats/index.shtml.

Library of Congress. “PDF/A-3, PDF for Long-term Preservation, Use of ISO 32000-1, With Embedded Files.” *Sustainability of Digital Formats*. http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml (last accessed 11-16-2012).

Littman, J. “BagIt for data managers.” https://docs.google.com/presentation/d/1FcFqLa3OUUA7uhYEbzQRfD3K5WLh2OTGYCE8e9fY36U/edit (last accessed 01-25-2013).

Lowe, D. and Bennett, M., (2009) “Digital Project Staff Survey of JPEG 2000 Implementation in Libraries.” *UConn Libraries Published Works.* Paper 16. http://digitalcommons.uconn.edu/libr_pubs/16 (last accessed 03-22-2013).

Microsoft. 2013. “Differences between the OpenDocument Spreadsheet (.ods) format and the Excel (.xlsx) format.” http://office.microsoft.com/en-gb/excel-help/differences-between-the-opendocument-spreadsheet-ods-format-and-the-excel-xlsx-format-HA010355787.aspx?CTT=5&origin=HA101878944.

Murray, K. 2013. “One Format Does Not Fit All: FADGI Audio-Visual Working Group’s Diverse Approaches to Format Guidance.” *The Signal*. http://blogs.loc.gov/digitalpreservation/2013/10/one-format-does-not-fit-all-fadgi-audio-visual-working-groups-diverse-approaches-to-format-guidance/ (last accessed 11-06-2013).

The National Archives. *(DROID) Digital Record Object Identification.* https://github.com/digital-preservation/droid#readme (last accessed 11-16-2012).

The National Archives. *PRONOM.* http://www.nationalarchives.gov.uk/PRONOM/ (last accessed 11-16-2012).

National Archives and Records Administration. 2011. “National Archives Digitization Tools Now on Github.” http://blogs.archives.gov/online-public-access/?p=6270 (last accessed 11-16-2012).

NDLTD. 2010. “ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations,” http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html (last accessed 11-30-2012).

MetaArchive Cooperative. 2012. “Integrity Checking.” *Preservation Planning Portal.* http://www.metaarchive.org/projects/neh (last accessed 11-16-2012).

MetaArchive Cooperative. 2012. “NEH Chronicles in Preservation.” http://www.metaarchive.org/projects/neh (last accessed 11-16-2012).

OCLC. *Persistent Uniform Resource Locators*. http://purl.oclc.org.

OpenDoc Society. *OpenDocument Format*. http://www.opendocumentformat.org/ (last accessed 11-16-2012).

Park, S. 2010. “HTML5 for ETDs.” *ETD 2010*. http://people.cs.vt.edu/~shpark/html5etds.html (last accessed 11-16-2012).

Rosenthal, D. S., Rosenthal, D.C., Miller, E., Adams, I., Storer, M., and Zadok, E. 2012.“The Economics of Long Term Digital Storage.” *LOCKSS* http://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf (last accessed 11-30-2012).

Taylor, N. 2012. “The Value of a Broken Link.” *The Signal.* http://blogs.loc.gov/digitalpreservation/2012/03/the-value-of-a-broken-link/ (last accessed 11-16-2012).

Todd, M. 2009. “File formats for Preservation.” *DPC Technology Watch Report 09-02.* http://www.dpconline.org/advice/technology-watch-reports (last accessed 01-25-2013).

VideoLAN Organization. “VLC media player.” http://www.videolan.org/vlc/index.html.

WebCite. http://www.webcitation.org/.

6 Metadata for ETD Lifecycle Management

Daniel Alemneh (University of North Texas)

Topics Covered

- Roles of metadata in facilitating the ETD lifecycle.
- Methods to capture metadata manually and automatically.
- Examples of programs using metadata to enhance ETD access.
- Strategies to manage metadata over time.

6.1 Introduction

Electronic theses and dissertations (ETDs) are an important output of the research cycle. Since the late 1990s, ETDs have played significant roles, not just as new forms of scholarly communication, but also as drivers for the development of institutional repositories and digital libraries in general. The successful management of ETDs requires effort throughout the entire lifecycle to ensure that ETDs are preserved and made accessible in a manner that today's users expect and that tomorrow’s users will find useful. As described in this guidance document, the creation and maintenance of information about the ETD files (including technical, functional, and descriptive metadata) is a key component of this effort.

Although the term “metadata” is used differently in different communities, metadata is usually defined as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO 2004). This document first identifies ETD metadata practices at different institutions and discusses the critical role of metadata in facilitating ETD lifecycle management. It then outlines some of the most important metadata elements to capture and discusses stakeholder roles and responsibilities in the creation of this information.

Some metadata elements are vital for ETD lifecycle management activities; others might be considered optional. Some metadata fields can be “extracted” from the files using software tools; other fields need to be filled out by hand. This document will recommend implementation strategies based on current ETD metadata best practices and standards.

6.2 Metadata Roles in ETD Lifecycle Management

Advancements in digital technologies shape the creation, access, use, and preservation of information resources in profound ways, including for electronic theses and dissertations. Although most ETDs still resemble their print equivalents and are text-based PDF files, supporting technologies have matured to allow for the creation and incorporation of complex and dynamic resources in an ETD.

In view of the central role played by metadata in digital libraries, the following section highlights some trends that impact ETDs’ curation as well as metadata best practices at large:

- Heterogeneous, multimedia ETDs instead of text documents.
- Complex retrieval systems instead of matching queries and document representations.
- Visualization of the information space instead of a ranked list of search results.
- Human information behavior instead of information needs.
- Access restrictions requested by the author, school, or research sponsor instead of applying a permanent status for the entire lifecycle.
- Users as both creators and consumers of information instead of one or the other.
- Long-term preservation of digital assets of unstable digital objects.
- Deployment of new cataloging standards in libraries, including RDA (Resource Description and Access) and the exposure of RDA-based data in the linked data cloud instead of traditional cataloging tools such as MARC (MAchine-Readable Cataloging) and AACR2 (Anglo-American Cataloguing Rules, 2nd revision).

6.2.1 Structural Metadata

Different disciplines have different ETD structures and requirements. Accordingly, institutional repositories typically support a variety of digital formats, as determined by the ETD program and documented in ETD policies. Candidates for a music doctoral degree, for example, may be required to submit a text-based dissertation, usually an analysis of a composition or a particular composer's works. This text document may also be accompanied by recitals that demonstrate the candidate's instrumental/vocal virtuosity and a lecture recital that combines a performance and description of the dissertation topic. Depending upon the policies of an institution, these may be recorded and stored as part of the ETD to provide additional evidence of the candidate's performance and fulfillment of academic candidacy requirements.

Accompanying materials in the ETD landscape are not limited to performances. Visual forms (geologic diagrams or maps, high resolution images of art objects, videos of lab observations, etc.) accompany dissertations in fields as diverse as environmental science, stratigraphic geology, art, and biology. ETD Programs and policies have taken different approaches to these additional objects – some separate the accompanying materials from the text of the ETD and some do not accept them at all. Increasingly, metadata is used as a way to bind these resources together, allowing academic institutions to provide and preserve multiple-format ETDs that include content such as video, audio, and datasets (Beagrie and Pink 2012).

Although institutions often require electronic theses or dissertations for graduation, some accompanying materials remain in their physical formats. Over the next decade, more theses and dissertations will be digitized retrospectively. Ensuring digital access to accompanying materials will be problematic for those accustomed to a single format and/or the digital representation of complete content. By employing appropriate metadata elements to record and link various characteristics and relationships, institutions can integrate ETDs' associated contents.

As described in the *Managing the Lifecycle of ETDs: Curatorial Decisions and Practices*, such integrations of complex content objects have implications for how ETDs are curated. A review of the current landscape in digital libraries and emerging trends shows that there is no shortage of opinions on the role of metadata in digital resources lifecycle management (Day 2006; Lavoie 2004). Various emerging Web applications – driven by semantic web technologies such as the Web ontology language (OWL), the resource description framework (RDF), semantic Web rules language (SWRL), and other members of the World Wide Web Consortium (W3C) family of specifications – offer powerful data organization, combination, and query capabilities.

6.2.2 Supporting Access Restrictions and Embargos

Limiting access to ETDs according to a student's campus, either geographically or by user identification, is a key intellectual property concern of ETD programs, and it is addressed with metadata. As stated in the *Guide to Access Levels and Embargoes of ETDs*, different institutions may have different access restriction polices that can be applied to all or part of a work. However, there are different options for handling restricted ETDs. Most institutions usually have a delay (1-5 years) and then the restricted ETDs move to public access. Depending upon university policies and student requests, the ETD and/or its metadata may be restricted.

To facilitate this, the metadata should contain information about why and for how long the ETD is restricted or embargoed. If the metadata is also restricted and/or embargoed, this should be coded in the metadata too. If this information is recorded in a standard way, it can be used by repository software to determine the status of an individual object and to change that status over time: e.g., if an ETD is embargoed for three years, the metadata should include information sufficient to allow a repository system to know the date upon which the embargo is lifted.

6.2.3 Facilitating Intellectual Property Rights Management

As discussed in the *Briefing on Copyright and Fair Use Issues in ETDs*, because ETDs capture the research efforts of students, higher education institutions have a responsibility to provide the best possible guidance on students' intellectual property rights. Among many stakeholders in the ETD lifecycle, libraries play an active role in the long-term stewardship of these resources. While students hold the copyright to their ETDs, they generally grant a library the right to preserve and provide ongoing access to their ETDs.

In light of the cross-disciplinary and cross-institution usage of ETD research, expressing rights and policy statements via metadata is vital for facilitating distribution of ETDs via institutional repositories. However, as stated in *Briefing on Copyright and Fair Use Issues in ETDs*, issues of copyright and intellectual ownership have been identified as serious concerns for most universities. Authors normally hold the intellectual property rights in the content of their ETDs. It is important to document the minimum core rights information that a repository must know, and what rights or permissions it has in order to carry out actions related to objects within the repository. These may be generally granted by copyright law, by statute, or by a license agreement with the rights holder. US copyright law, for example, governs all items published since 1923, which may differ from other countries. Knowing the birth and death dates of the creator and the year in which the ETD was created will help to calculate and determine the copyright status. For example, if the creator designated his/her estate as the copyright owner upon his/her death, performance and use of the compositions is the right of the composer and his/her estate until seventy-five years after the composer's death.

Moreover, in light of transitional events in ETDs' lifecycle (embargo releases, redactions, and other possible preservation actions), rights metadata information needs to track and document the changes continually. New rights information may be provided or discovered by the copyright owner, users, or other parties. Also, new legislation and policies at national and/or institutional levels will likely require changes in rights metadata information in order to reflect the most current right status of the ETDs.

6.2.4 Facilitating Preservation Activities

Recognizing the critical role of metadata in any successful digital lifecycle management strategy, institutions that take responsibility for digital objects should also implement a metadata-based approach to ensuring long-term access. In addition to the descriptive metadata, (which describes the intellectual entity and supports discovery and delivery of ETD content), preservation metadata provides provenance information, documents preservation action, identifies technical information, and helps in verifying the authenticity of digital objects.

Essentially, the ETD metadata needed to support the full range of digital preservation activities can be loosely categorized into the following possibly overlapping groups:

- Provenance metadata: records the origin or provides an historic context or source provenance, such as specifying the analog source material for a digital derivative.
- Structural metadata: captures physical structural relationships, such as which image is embedded within which file, as well as logical structural relationships, such as page order, in born digital or digitized ETDs.
- Technical metadata: captures format-specific technical information that applies to any file type, including information about the software and hardware on which the digital object can be rendered or executed, as well as checksums and digital signatures to ensure fixity and authenticity.
- Administrative metadata: provides provenance information regarding who has cared for the digital object and what preservation actions have been performed on it. It also provides rights and permission information that specifies embargoes and access of ETDs and which preservation actions are permissible. As discussed in *Briefing on Copyright and Fair Use Issues in ETDs*, rights take different forms and some rights are covered by laws, like copyright law, which can be different in different jurisdictions. There are also contractual rights (e.g. licenses) that are governed by contracts (terms and conditions) between parties.

Although there is no "one-size-fits-all" solution to the digital preservation problem, the role of metadata in ensuring long-term access and management has been analyzed by many researchers. The PREMIS (PREservation Metadata Implementation Strategies) working group holds Preservation Health Check workshops and implementation fairs on a regular basis with real life examples of preservation metadata. Such a forum can help early adopters, practitioners, and other stakeholders to come together and refine theoretical assumptions in order to build a better shared understanding of what, why, and how preservation metadata are collected and created (PREMIS 2014).

6.2.5 Describing ETDs

Because dissertations must constitute original research, each is unique to the bibliographic world. As such, each dissertation receives original cataloging and metadata descriptions (Alemneh and Hartsock 2014). In many cases users want to know what has been done in a particular discipline or topic area (Voorbij 1998). Although consulting the bibliographic detail or abstract alone is inadequate for scholarship, the complete metadata description can convey sufficient information about ETDs to give users confidence that the ETD will be relevant to their research areas.

Metadata elements describe characteristics or aspects of an object or digital resource. With enhanced metadata-based and subject-specific search mechanisms, it is now easier than ever to access, use, and reuse scholarly works and associated data that have not been available through traditional publishing methods.

Assigning appropriate metadata to ETDs can improve discoverability by increasing their visibility. To describe digital resources accurately, metadata creators try to follow, as closely as possible, the thinking of the creator/author and also to anticipate what users might want to discover and how they will retrieve the information. As noted by many researchers, one of the key issues for information retrieval and all other content-based text management applications is document indexing (Cleveland and Cleveland 2013). The generation of accurate indexing terms is fundamental to the discovery, use, and reuse of digital resources. An index term is simply a systematic representation of an information-bearing object (text, image, audio, video, etc.) that points users to specific items on topics of interest. In other words, it is an information retrieval tool.

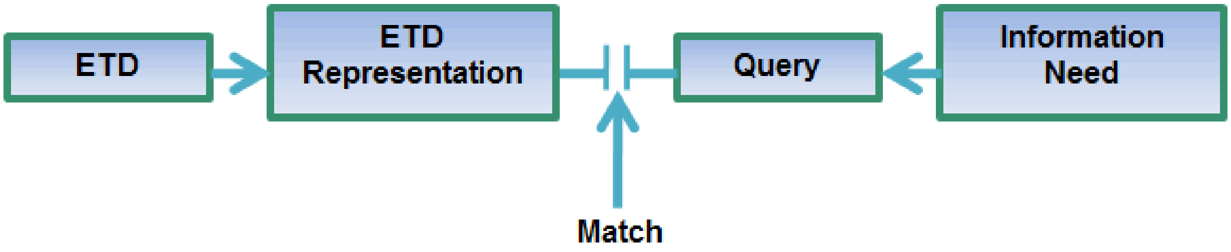


Figure 6-1. The classic Information Retrieval model (modified from Bates 1989)

Indexing enhances the accessibility and value of a resource, provided that it is based on a thorough analysis of that resource. Similarly, subject metadata conveys what an ETD is about. In addition to the needs and abilities of the searchers, the subject terms needed to sufficiently create access to ETDs depend on the content that is indexed. A good index helps users find what they need, even when they are not sure of what they are looking for. Lancaster (2003) noted that in the rapidly growing information environment, unidentified and disorganized content, however useful it might be, is at risk of being rendered undiscoverable, and thus obsolete. *Figure 6-1* depicts the most common forms of document representation that are familiar from the classic model of information retrieval.

Many information retrieval researchers agree that in light of the continual evolution of users' information seeking behavior systems should be sufficiently flexible to allow users to adapt to the current environment (Bates 1998; Ellis 2004; and Kuhlthau 2006). Search engines generate results by matching search terms entered with terms in the content, usually referred to as full-text searching. "Among young academic researchers in Sweden, Google was the most-used starting point for searching scientific information" (Haglund and Olsson 2010). This form of searching has shortcomings, mainly in precision. Precision refers to the proportion of the relevant documents retrieved in a search to the total number of documents retrieved in a search. Increasingly, search engines use metadata to enhance precision and improve their search results. Similarly, many researchers noted that in comparison to ten years ago the use of metadata by Google has increased considerably (Beall 2010).

In contrast, many information scientists still argue that we have not developed comparably mature tools for exploratory search – information seeking where users do not have a known target document and may not even have a well-established information need. "Only in the last few years have we seen an emerging program of human–computer information retrieval (HCIR) that brings interactive techniques to bear on more sophisticated information seeking tasks" (Ardo 2011).

6.3 Capturing Metadata

Assigning appropriate metadata to ETDs can improve discoverability by increasing their visibility (Ivanovic et al. 2012). There are data elements that serve a function both for the users' tasks of search and retrieval (bibliographic records) and for digital preservation purposes (preservation metadata). Users are more interested in the content and the subjects, than in what format the objects are delivered. Many commentators agree that the most useful metadata about a digital object are the subjects or keywords, since they explicitly describe what the object is about (Schwing et al. 2012). A number of researchers (Peterson 2010; Spiteri 2011) have analyzed content indexing (especially subject indexing) and described the general information seeking behavior of users. Many agreed that the two major reasons why users experience problems with subject access are the quality and application of subject indexing on the one hand, and the complexity of users' knowledge and information literacy skills required for successful subject access on the other. To maintain the consistency of search results and high recall of available resources, it is critical to ensure the quality of the keywords and taxonomies used to index heterogeneous digital resources within digital libraries.

6.3.1 Metadata by Librarians

Catalogers and information professionals have always done some assessment of material. To fully understand what a good index is, it is necessary to be both micro- and macro-minded. On the micro level, we concern ourselves with the specific mechanics of creating an index term. On the macro level, indexing is part of the larger context of an information retrieval system. Retrieval of information involves the user expressing an information retrieval request that incorporates terms from the common vocabulary to match stored records.

In this regard, controlled terms provide a broad navigational tool for browsing through digital content and digital library collections. Controlled vocabulary terms ensure indexing consistency and enhance retrieval precision across all digital resources. In many institutions, catalogers adapt existing workflows and procedures to handle ETDs. The creation of digital cataloging and metadata workflows is a good opportunity to implement general cataloging policy changes.

One response is outsourcing some segment of metadata/bibliographic processing or related technical service operations. This strategy is increasingly seen as a way to cut technical service costs and enhance efficiency. However, outsourcing requires investment in quality control, potentially taking the form of experienced or professional librarians to monitor output.

Another response is opening new positions that facilitate metadata creation: scholarly communication librarian, repository librarian, digital curator, copyright librarian, semantic technologies librarian, etc. Regardless of their labels or names, the primary and common responsibilities of such positions include formulating plans for moving their libraries forward to meet the challenges of changing modes of scholarly communication. This usually includes an active role in promoting open access and developing a metadata strategy for the library and communicating the implications of such strategies to the university community. Their liaison with faculty and students provides opportunities to integrate digital contents and digital repositories into the learning and research mission of their institutions.

6.3.2 Author-Supplied Metadata

In addition to metadata created by professionals, incorporating author-supplied keywords can enhance descriptive metadata. Some libraries (including early ETD adopters such as Virginia Tech) incorporate only author-supplied keywords. However, considering the growing interdisciplinary nature of higher education and the importance of topical approaches for ETD users, the subject matter may not be sufficiently captured by authors' supplied terms alone.

In view of the constant changes in users' requirements, access to ETDs relies on a seamless discovery process that offers multiple options to users. Incorporating both controlled and natural approaches to the subject matter in an ETD collection, results in high-level descriptions and representations. Alemneh and Rorissa 2012 noted that such approaches (including using author supplied terms and social tags as possible sources of metadata) add value and enhance user's ability to find, access, use, and re-use digital objects.

6.3.3 Automatic Extraction

There are a growing number of metadata extraction tools that enable works to be automated either through batch processing or processing on an individual basis as required. Automated metadata extraction is particularly useful for collections like ETDs. For example, the National Library of New Zealand's Metadata Extraction Tool can programmatically extract preservation metadata from a range of file formats including PDF documents, image files, sound files, Microsoft Office documents, and many others. Although most such tools are designed to facilitate self-documenting and preservation activities over time, they can also be used for other tasks such as the extraction of metadata for resource discovery.

In addition to technical metadata, there are many other vital ETD characteristics that need description, tracking, and integration. Researcher profiles (or author/contributor metadata) can often be collected from an institution's human resources system (e.g., Banner, PeopleSoft). In most institutions, these systems are tied to the author's plan of study and can supply additional information such as academic discipline, committee members, and document type.

There is growing consensus among academic and research communities about the critical need to link publications and underlying data. In this regard, the current customized version of DSpace-CRIS (Current Research Information System) can be cited as one of the emerging new add-ons that enable the ingestion, storage, display, and management of metadata and full text for ETDs and other research entities. Using such tools, items from native system and/or new resources from different systems, can be linked to each other using auto-complete and auto-lookup functions in the submission edit phase.

Considering the emerging multi-format/part structures of ETDs, tools that simplify a smooth integration between local resources and other entities are particularly useful. Quality metadata plays a significant role in integrating and contextualizing all of these heterogeneous entities. In turn, this adds greater value to each individual component and facilitates visibility, discovery, and understanding of the overall research agenda.

6.3.4 User-Generated Terms – Post-Ingest Activities

Although different metadata schemas are often complementary, good keyword terms help users find what they need, even when they are not aware of their needs. In the quest for better discoverability of existing digital resources, libraries implement systems to enrich traditional catalog and metadata records with additional user-supplied terms and descriptions.

A growing number of institutions are assessing the potential for user-supplied tags or folksonomy to complement established controlled vocabularies in a diverse and collaborative environment. A folksonomy is any system that allows users to tag their favorite digital resources using natural language words. In a socially constructed metadata paradigm, users not only search/browse, access, and use content but also proactively participate in its production and description by tagging, rating, reviewing, highlighting, and recommending (Alemneh and Rorissa 2014; Smith 2008).

With the increasing popularity of social tagging systems, Sue Yeon Syn and Michael B. Spring (2013) explored and analyzed social tagging systems as a mechanism to allow nonprofessional catalogers to participate in metadata generation. The results suggest that user tags successfully identify the terms that represent the topic categories of web resource content. However, user-generated terms have been suggested as a lightweight way of enhancing descriptions of digital contents and improving their access through broader indexing. Trant (2008) summarized both the negatives and positives of folksonomies. On one hand, critics point to the fact that it is an uncontrolled vocabulary and leads to less effective information retrieval. On the other hand, proponents of the concept of crowd sourcing point to the fact that it is user-friendly and enables personalized information retrieval by users.

As folksonomies are in a continual state of flux, they are better able to accommodate current terminology and concepts than traditional indexing tools and systems such as the Dewey Decimal Classification and the Library of Congress Subject Headings. Traditional approaches share a basic problem: the potential users of information are disconnected from the process used to describe that information.

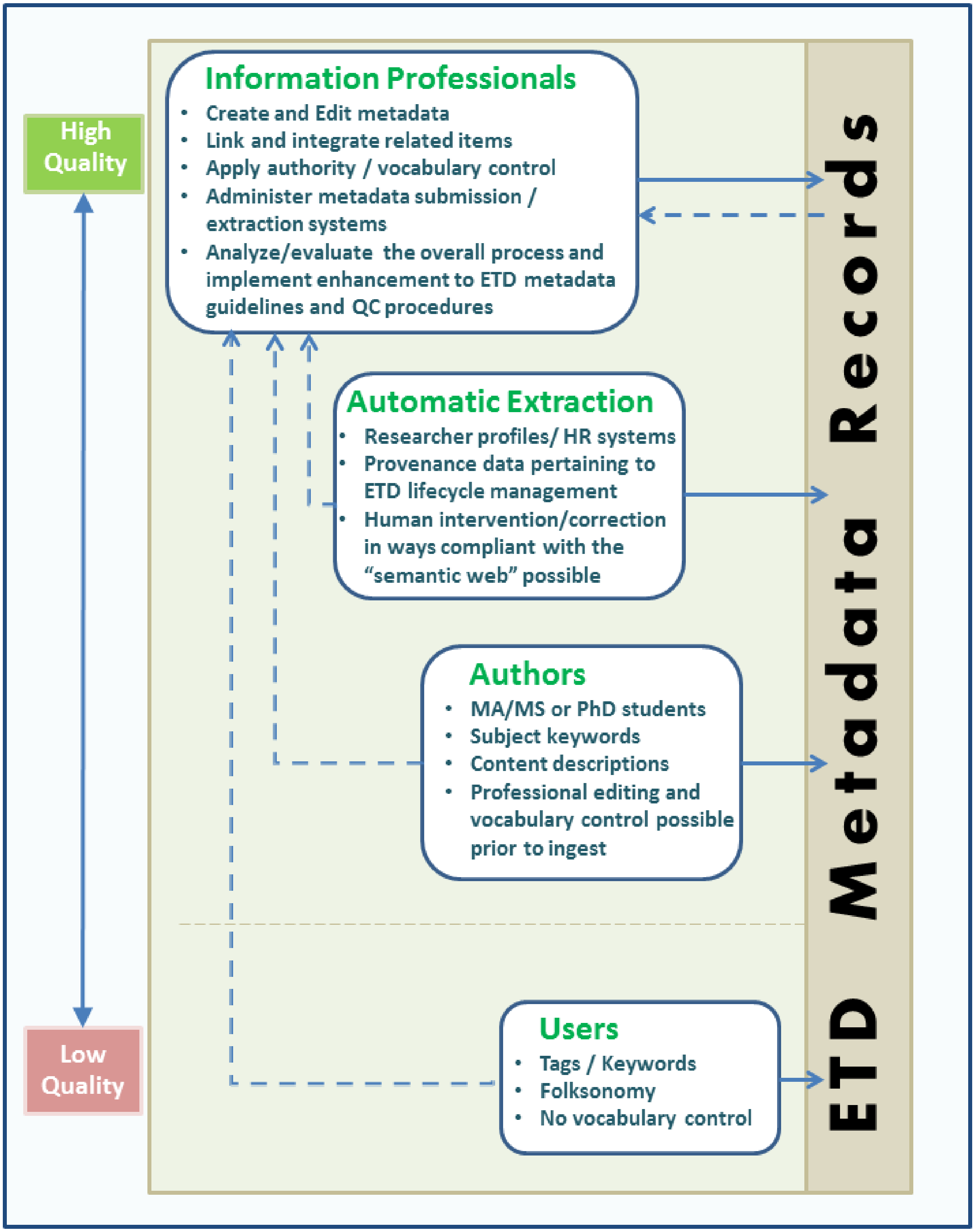

High Quality
Low Quality
Information Professionals
Create and Edit metadata
Link and integrate related items
Apply authority / vocabulary control
Administer metadata submission / extraction systems
Analyze/evaluate the overall process and implement enhancement to ETD metadata guidelines and QC procedures
Automatic Extraction
Researcher profiles/ HR systems
Provenance data pertaining to ETD lifecycle management
Human intervention/correction in ways compliant with the "semantic web" possible
Authors
MA/MS or PhD students
Subject keywords
Content descriptions
Professional editing and vocabulary control possible prior to ingest
Users
Tags / Keywords
Folksonomy
No vocabulary control
ETD Metadata Records

Figure 6-2. ETD metadata creation by different stakeholders

Combining traditional indexing systems with folksonomies is the solution for delivering a richer digital library user experience. However, socially constructed metadata approaches have not been effectively implemented so as to augment legacy library metadata. The problem could partially be attributed to the absence of a conceptual metadata framework that could serve as a theoretical basis for a better understanding of the possible uses of Web 2.0 services in libraries (Lagoze 2010). There are small but growing initiatives in various libraries that have effectively engaged the user community from transcribing content to correcting OCR output. Tapping user knowledge does not mean that the information professionals curating a collection in a production environment would be able to step back from the collection after the core metadata was added. As described by many commentators and depicted in *Figure 6-2*, involving all possible stakeholders and building a dedicated group of contributors requires an investment of repository staff time, as well as strategies for promoting data ownership, ensuring quality, and producing seamless integration.

6.4 Best Practices for ETD Metadata

Maintaining usable ETDs requires high-quality metadata about those digital objects. An effective metadata management approach improves consistency, clarity of data lineage, and relationships between and among objects so that institutions can better integrate resources. There are many metadata practices that are implemented by the ETD community at national, regional, and even international levels. A definition of what constitutes minimal, good, and optimal metadata – regardless of format or schema – depends on many factors. These factors can differ from country to country and from institution to institution.

Quality metadata plays a significant role in facilitating the establishment of a union catalog, which is an indispensable element of library networking and resource sharing. Some countries coordinate activities related to ETDs and work with institutions of higher education to ensure high quality metadata. While ETDs are maintained by the institutions at which they were produced, it is possible to give searchers the appearance of a single collection by gathering all the metadata into a central search engine. When a potentially relevant document is found, the systems redirect the user to the institution that houses the document.

A number of institutions and consortia have developed or instituted local metadata best practices. The following sections provide further description. Although the adopting institutions had already validated the practices as “best” by their standards, they may not be suitable for every institution.

6.4.1 International Initiatives

6.4.1.1 Networked Digital Library of Theses and Dissertations (NDLTD)

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of ETDS. Since its inception in 1996, NDLTD has worked to improve graduate education, increase the availability of student research, empower students and universities, advance digital library technology, and lower the costs of submitting and handling ETDs.

NDLTD has developed a metadata standard especially suited to ETDs (see *Table 6-1*). NDLTD provided the following metadata elements as a guideline to develop a faithful crosswalk between local metadata standards and a single standard used for sharing information about ETDs.

No.	ETD Metadata Elements MARC Field & Subfield		MARC Field & Subfield	Description
	Top-Level	Sub-Level		
1	title		245a	*Full title for the work as it appears on the title page.*
		Alternative title	246	*For title, choose the appropriate title qualifier, (main, alternative, added, translated, etc.), preferably from the controlled list. (UNT Libraries, "Title Role Authorities")*
		(Title qualifier)		
2	creator		100a	*Name of Author*
3	contributor		720a	*Name of Committee Member, Thesis Advisor, chair*
		(Contributor role qualifier)		*For Creator / contributor roles, choose the appropriate qualifier, (Committee Member, Thesis advisor, Examiner, etc.), preferably from the controlled list.*
4	subject / keywords		653a	*Subjects and/or keywords describe what the ETD is 'about' and enter as many terms as necessary to capture subject content.*
5	description		520a	*Content and Physical Descriptions*
		abstract	520a	*Abstract usually supplied by ETD authors)*
		note	5xx (500, 504)	*Only include notes when applicable. (Example release note.)*
6	publisher		260a+b	*Usually the institution's name*
7	date		260c	*(publication date, graduation date,)*
8	type		Leader 6&7	*Type of resources*
9	format		856q	*Physical Description, (preferably from the controlled list)*
10	identifier		856u	*Unique identifiers include URL of an ETD.*
11	language		008	*If the ETD is in multiple languages, include each of them, preferably from the controlled list.*
12	coverage		651 or 690	*Coverage information (usually place and time period)*
13	rights		540	*Information about rights may include:* **Access (level of access that will be allowed to users)* **License (if there is a license or rights)* **Holder (usually ETD author is the rights holder)* **Statement (information about:* *-Rights held in and over the resource,* *-Conditions under which the work may be used, distributed, reproduced, etc.,* *-Information about how the rights conditions may change over time,* *-Whom to contact regarding the copyright of the work.)*
14	Degree		502a	*Degree information usually includes degree name, level, discipline, and name of institution and academic department granting the degree associated with the work.*
		Degree name	502a	
		Degree level	502a	
		Degree discipline	710b	
		Degree grantor	502a; 710ab	
15	*Other**			*Other optional descriptive metadata elements can be used or added as appropriate:* **Relation (can be used when related items are online such as a doctoral recital and the ETD connected to it.* **Metadata Information (who creates/updates what, when, etc.)*

Table 6-1. Modified NDLTD's descriptive metadata for ETDs with semantic MARC crosswalk.

The NDLTD metadata is intended to be flexible enough to be used in a variety of current and future representations of ETDs. If feasible, as suggested in *Table 6-1*, compiling local controlled vocabularies or assigning metadata values from controlled lists facilitates consistency. Whenever possible, while accommodating local requirements, using existing or widely adopted controlled vocabularies promotes even greater interoperability.

The NDLTD metadata has been used by institutions and partners around the world to export their public metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). As noted by Ivanovic, Ivanovic, and Surla (2012), visibility of ETDs can be increased by putting the digital object or its descriptive metadata (or both) into such networked repositories. Although the NDLTD union archive now contains more than two million records of ETDs, based on the total number of ETDs produced around the world (estimated to be close to 1 million per year), it is far from being a comprehensive union catalog for global ETDs.

6.4.1.2 PREservation Metadata Implementation Strategies (PREMIS)

Addressing the preservation and long-term access issues for digital resources is a significant challenge for repositories. A number of researchers noted that the problem of ensuring long-term access to digital information sources is compounded by the fact that most of the sources are not properly organized.

Most agree that extensive metadata is the best approach to minimize the risk of digital objects becoming inaccessible (Alemneh and Hastings 2010).

Accordingly, a number of national and international projects and initiatives attempted to assess the potential role of metadata in preservation management activities. PREservation Metadata Implementation Strategies (PREMIS) has been influential in providing a "core" set of preservation metadata elements that support the digital preservation process.

PREMIS is a core set of metadata elements (called “semantic units”) recommended for use in all preservation repositories regardless of the type materials archived, the type of institution, and the preservation strategies employed (PREMIS 2014). The PREMIS data model consists of five interrelated entities: intellectual, object, event, agent, and rights with each semantic unit mapped to one of these areas:

- Intellectual entities are conceptual and might be called "bibliographic entities." PREMIS does not actually define any metadata pertaining to intellectual entities as there are plenty of metadata standards to choose from, for example NDLTD Descriptive Metadata for ETDs.
- Objects are what are actually stored and managed in the preservation repository. Although descriptive metadata is out of scope, objects can point to descriptions of intellectual entities or the entity itself, in either direction.
- The event entity aggregates information about actions that affect objects in the repository. Events are extremely important for preservation activities, as it is vital to track actions performed on digital objects (such as capture, compression, validation, replication, migration, etc.) for preserving into the future.

- Agents are actors (people, organizations or software) that have roles in events and in rights statements.
- The rights entity aggregates information about rights and permissions that are directly relevant to preserving objects in the repository (PREMIS 2014).

As can be seen in the diagram of PREMIS data model relationships (see *Figure 6-3*), "objects" can have relationships to other entities in the data model or to each other. However, "agents" can only act on objects through events or rights, not directly. *Table 6-2* lists the elements that are properties of the given entity. Most of PREMIS is actually devoted to describing digital objects that include technical metadata. In addition to format and size, it also includes things like fixity, commonly known as "checksums," and any passwords or encryption that limits the use of an object. Environment details the software and hardware needed to render or use the object. Relationships are classified as structural, how objects comprise the original ETD relate to one another, or derivative, how service or access copies relate to the original ETD.

Although all of the five entities are interrelated, they can be used and implemented independently from each other. Institutions may adopt one or more PREMIS entities. Decisions on how to implement the recommended entities remains entirely up to the individual repository. This allows a repository to implement localized ETD workflows and submission models. Many academic institutions have implemented object, event, and agent entities. For example, the University of North Texas Libraries has built a tool for capturing and providing user access to PREMIS events that are important in a digital objects lifecycle. Such a tool is fairly focused on a part of the PREMIS model (event and agent) and solves a specific problem, which is to capture events in an objects lifecycle (ingest, fixity check, virus check, replication, migration).

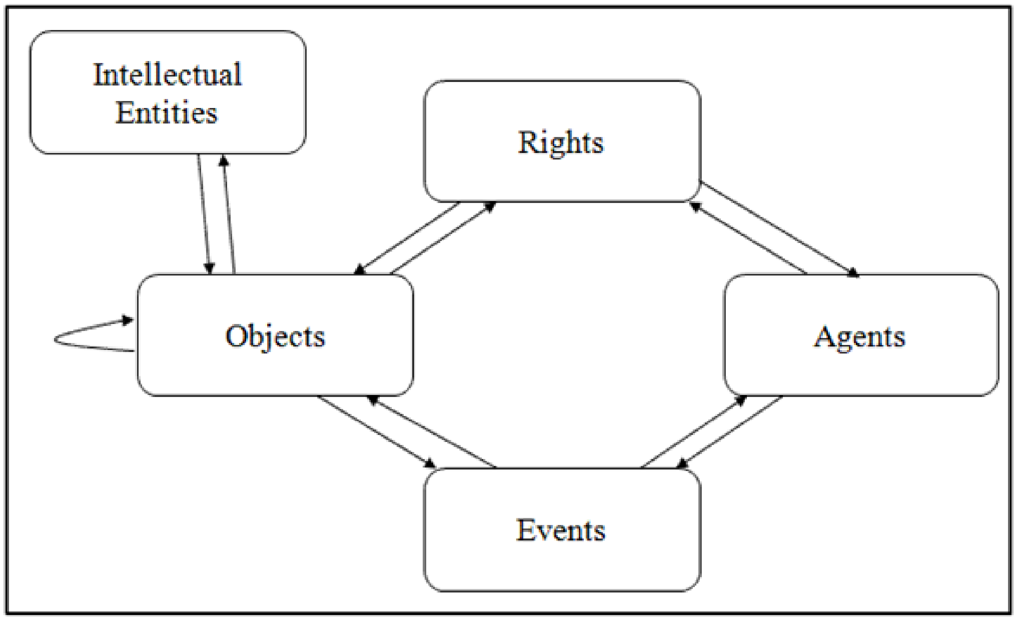


Figure 6-3. The PREMIS data model – Version 2.2 (PREMIS 2014)

Object ***(Information about the stored digital objects)***	**Event** ***(Information about actions)***	**Agent** ***(Actors that have roles)***	**Rights** ***(Rights information)***
Object ID *(type and value)*	Event ID	Agent ID	Rights statement *(ID: type & value)*
Preservation level	Event type	Agent name	Granting Agent
Object characteristics *(format, size, fixity, etc.)*	Event date/time	Designation *(person, software, organization).*	Permission granted
Storage	Event outcomes		
Environment *(Hardware, software)*	Linking identifiers		
Digital signatures			
Relationships			
Linking identifiers			

Table 6-2. Summary of types of information included in PREMIS by entity type (PREMIS 2014)

The rights entity is of particular interest to the ETDs community. Access rights may be generally granted by copyright law, by statute, or by a license agreement with the rights holder. Repository rights only concern rights to preserve, and an institutional repository needs to know what permissions it has to carry out appropriate preservation actions related to ETDs within the repository. For the purpose of the PREMIS Data Dictionary, statements of rights and permissions are encompassed by the rights entity:

- Rights are entitlements allowed to agents by copyright or other intellectual property law.
- Permissions are powers or privileges granted by agreement between a rights holder and another party or parties.

A number of early PREMIS adopters realized that the rights entity lacked the robustness required by various types of digital objects including ETDs. To address such limitations, the PREMIS Editorial Committee has been working on changes and enhancements to the rights entity semantic units. The PREMIS Editorial Committee incorporated several requested changes to the rights entity and published the greatly expanded version of the PREMIS Data Dictionary for Preservation Metadata version 2.2 in May 2012. Based on several other requests from early implementers, the PREMIS Editorial Committee is also working on a major revision for version 3.0 to include rights pertaining to a license, copyright and statutory rights. It is expected that the majority of this information will be captured from the object itself or from the repository system being used and that a minimal amount will need to be provided by hand (PREMIS 2014).

6.4.1.3 ProQuest Theses and Dissertations

ProQuest Theses and Dissertations (PQDT), is a database of dissertations and theses, published electronically or in print. PQDT includes nearly 3 million searchable citations to dissertation and theses from around the world from 1743 to the present day. More than 80,000 new full-text dissertations and theses are added to the database each year through dissertations publishing partnerships with 700 academic institutions worldwide. Although most ProQuest services are available for purchase, ProQuest has a DTD (Document Type Definition) that describes the XML (Extensible Markup Language) feed delivered to universities for free (ProQuest 2014).

6.4.1.4 Statewide Initiatives

A number of regional associations in the US such as the Ohio Electronic Thesis and Dissertation Association (OETDA), the Texas ETD Association (TxETDA), and the Florida ETD Association (FLETDA) seek to increase ETD knowledge of their members through the sharing of state-wide, national, and international best practices. Likewise, regional ETD consortia such as Texas Digital Library (TDL), OhioLINK ETD Center, and State of Florida ETD System, provide resources and services to their member institutions. These associations and organizations also provide networking and learning opportunities, establish standards, promote ETD open access, and more.

The Texas Digital Library (TDL), for example, is a consortium of higher education institutions in Texas that provides shared services in support of research and teaching. To facilitate ETD workflows, the TDL created the Thesis and Dissertation Submittal System (TDSS), more commonly known as Vireo. The Metadata Working Group of the TDL has also developed a descriptive application profile for ETDs in the Metadata Object Description Schema (MODS). Though other metadata formats were discussed and considered, MODS was chosen primarily because it was based on MARC, and it could capture descriptive information (which is analogous to the traditional catalog record) more easily than other schemas, such as Dublin Core. The only extension needed was degree information. With MODS as the descriptive metadata for ETDs, the working group created a TDL ETD MODS application profile. The schema defines 16 top-level elements that are included in the Vireo ETD Submission System and adopted by most TDL members' institutions (see *Appendix A: Texas Digital Libraries (TDL) ETD Metadata Example*).

6.4.1.5 Non-US ETD Initiatives

There are several ETDs projects outside of the United States. One of the objectives of an ETD program is to provide easy global access to the results presented in ETDs, irrespective of the languages or where the ETD was written. In this regard more and more international professional societies and organizations, such as ASIS&T (Association for Information Science & Technology), are expanding their dissertation awards for dissertations in other languages, explicitly stating that dissertations are welcome in any language. Nevertheless, implementing widely adopted metadata using English in addition to the original languages (e.g. translating titles and abstracts) facilitates and enhances international access.

Among other successful ETD best practice encouragers, the European ETDs collaborative projects can be mentioned as one of the many international efforts that identified metadata for use when describing the content of an ETD repository. The Dublin Core-based recommendation was the result of collaboration between UK institutions that have been involved in developmental work associated with ETDs. There are a number of other initiatives that aim to develop national and regional resource discovery services and search tools to promote the visibility of ETDs at individual member sites. The following four can be cited as success stories, in terms of promoting ETD-specific best practices:

- Cybertesis: A collaborative program (among 32 universities of Europe, Africa and Latin America) which is a portal developed jointly by the University of Chile, the Universites de Lyon, UnMontreal, and Alexandrie, and the University of Geneva for accessing full text ETDs from many countries, including Bolivia, Brazil, Canada, Chile, France, Hong Kong, Mexico, Peru, Spain, and the United States (Global Open Access Portal 2014).
- Database of African Theses and Dissertations (DATAD): The DATAD database contains citations and abstracts for theses and dissertations completed in African universities. Although African theses and dissertations contain local empirical data that is not available in international literature, African research results are rarely indexed in major international databases. In 2000, the Association of African Universities (AAU) initiated and supported efforts toward putting Africa's research output onto the mainstream of world knowledge via DATAD (AAU 2014).
- UK's Electronic Theses Online Service (EThOS): Offers a single point of access to all UK theses. EThOS harvests e-theses from institutional repositories and digitizes paper theses free of charge (EThOS 2014).
- Theses Canada: A union catalog of Canadian theses and dissertations, in both electronic and analog formats. Canadian universities participate in the program voluntarily by submitting approved theses and dissertations to Theses Canada (Library and Archives Canada 2014).

Such collaborative national and regional projects support the open access aspirations of many institutions and provide visibility to the work of scholars both within and outside of the countries of origin. Some initiatives (such as EThOS) help institutions to digitize analog copies and return the digitized versions to local institutions for loading onto their respective institutional repositories. More importantly, these organizations coordinate various activities – including the compilation of national and regional union catalogs for ETDs – and enable the participation of many partner institutions, small or large, with or without an institutional repository.

All these collaborative activities are facilitated by means of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which is a protocol for exposure of metadata rather than content. Using the OAI-PMH, individual sites can make their metadata accessible to search providers and discovery services and still maintain complete control over the resources. Supporting the OAI-PMH at individual institutions is the best way to contribute to such endeavors and promote repository interoperability.

Many institutions provide public access to a number of application programming interfaces (APIs) to their ETDs collections that can be used openly by those interested in programmatically accessing data from such systems. A growing number of discovery services, such as Open Access Theses and Dissertations (OATD), work with institutions' metadata to index only records that are actually ETDs and are freely available online. In addition to harvested metadata, they include a growing amount of full text "snippets" (about the first 30 pages of the thesis and sample images) to help guide searchers.

6.5 ETD Metadata Quality and Management

As demonstrated above, although ETDs are produced at individual institutions, various statewide, national, and international consortia play vital roles in developing best practices for ETDs management.

They also create environments for early ETD adopters to share and promote their best practices with other institutions pursuing ETDs as a catalyst for digital libraries development.

For the last two decades ETDs have played significant roles, not just as new forms of scholarly communication, but as drivers for the development of institutional repositories and digital libraries. In this regard, cultural heritage institutions are making good progress in digital preservation, specifically in the implementation of preservation metadata schema such as PREMIS. ETD-specific lifecycle management guidelines contribute to preserving and ensuring long-term access to digital resources.

Metadata management provides the tools, processes, and environment to enable repositories to answer a number of questions related to the digital object. An effective metadata management approach can

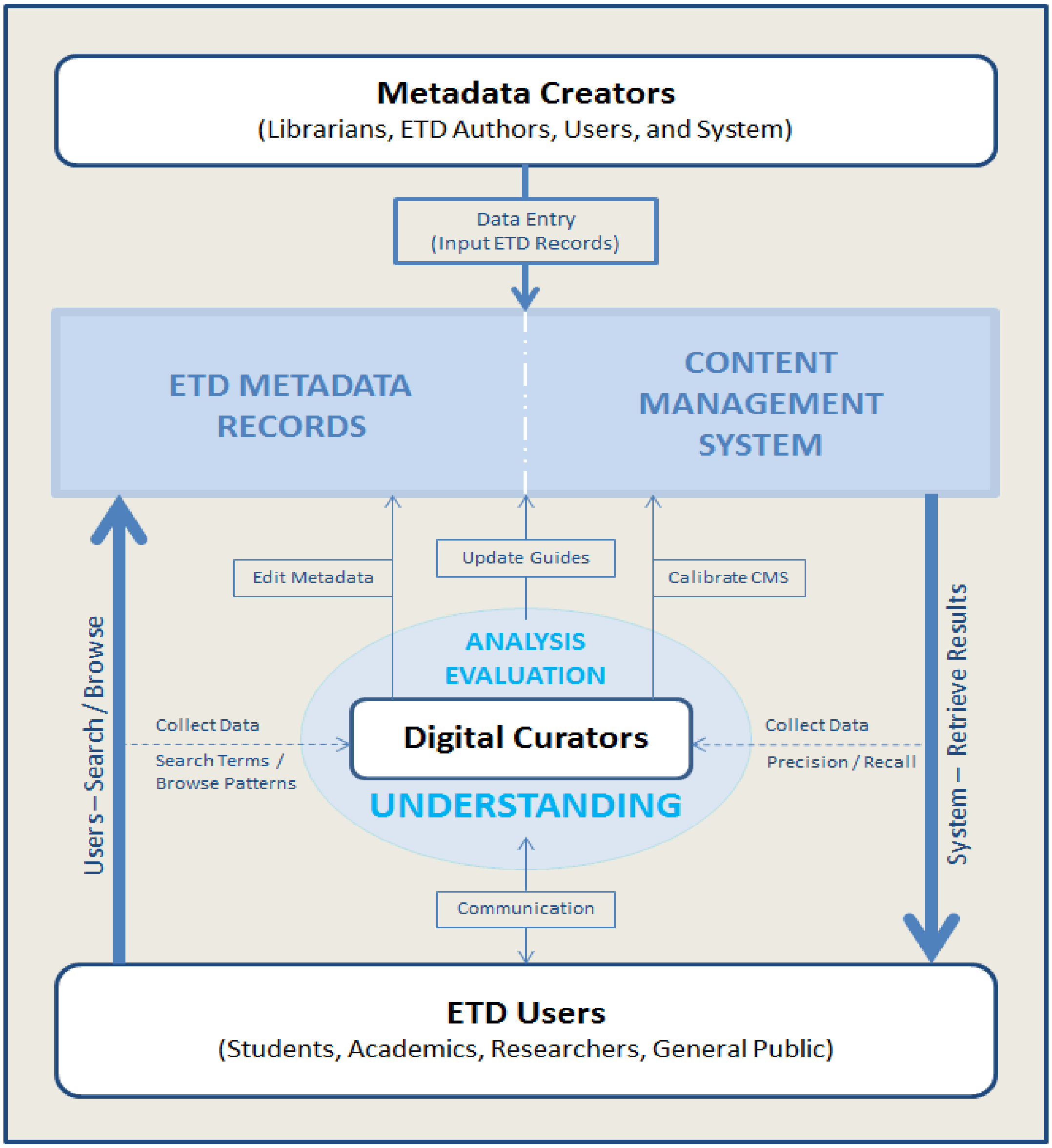


Figure 6-4. Flowchart for maintaining high quality ETD metadata

help institutions improve consistency, clarity of data lineage, and relationships so that they can better integrate resources. Integrating and contextualizing all of the ETDs components along with original research data and pre-and post-publications or performances, adds greater value to each individual component.

6.5.1 Metadata Quality

Maintaining usable ETDs requires high-quality metadata about those digital objects. Metadata quality characteristics depend on various factors, including: metadata record completeness, consistency, accuracy, provenance, conformance to expectations, and other local criteria and known substantive factors. Metadata quality issues are particularly acute if there are multiple institutions participating in collaborative ETD projects where a high level of interoperability is an important element.

Figure 6-4 shows the continuous metadata quality assurance loop. High levels of precision and recall, the two ways in which we measure any information retrieval system, are dependent on many activities that require the involvement of both machine and human interventions. In the figure, arrow line weights reflect the relative importance of each activity. Arrow labels describe physical information retrieval and system administration activities. The central oval represents intellectual activities performed by digital curators. In addition to employing workflows compliant with national and international standards, digital curators must be able to understand how users access ETDs, and then adjust the system for optimum retrieval.

6.6 Summary

The successful management of ETDs requires effort across the entire lifecycle to ensure that ETDs are preserved and made accessible in a manner that today's users expect and that will be useful to scholars in the future. There are several players over the entire ETDs lifecycle.

Considering the growing interdisciplinary trends in higher education and constant development of information and knowledge management, ETDs demand specialized treatment and characterization that can best capture the semantics and relations of the underlying concepts.

Most academic institutions are interested in making sure that their scholarly output is available to the widest possible global audience. Repositories employ a metadata based dissemination strategy in order to facilitate access, discovery, and use. In order to thoroughly describe ETDs and achieve the required data quality, it is important to engage stakeholders at all stages of the ETD lifecycle. Depending on the roles of the stakeholders, the types of information that should be captured for ETDs and by whom varies, not just at the point of ingestion, but also subsequently, as ETDs often have supplemental files (video, audio, data sets, errata, etc.) and transitional events in their lifecycle (embargo releases, redactions, etc.).

Integrating and contextualizing all ETD components and optimizing the content for search engines adds value and brings greater visibility to individual components across the repository. Most ETD repositories are OAI-PMH compliant, which means that their ETDs can be harvested into the OAI-based union catalog, regardless of the software platform. These multiple approaches enhance an ETD user's ability to find available resources while leveraging the benefits of composite applications, mash-ups, and service-oriented architectures.

In addition to the description of required metadata elements and the rationale for why they should be used or adopted, this document also narrates the process of creating the metadata necessary to provide complete access to the users of ETDs. Observations about the ETDs metadata, together with the best practices and their associated framework, are offered in the spirit that they may serve as an implementation roadmap for: creating shareable metadata, ensuring interoperability, and assisting the

ETD community in meeting the larger digital curation and lifecycle management challenges. Such practice will have important implications for future researchers, potential ETDs users, and stakeholders engaged in various aspects of digital curation efforts in general.

Here it should be emphasized that the implication of ETD lifecycle management will not be limited to the ETD community or higher learning institutions. Stakeholders benefit from metadata management ranging from the university community to external partners in business, industry, government, and society at large.

Appendix A: Texas Digital Libraries (TDL) ETD Metadata Example

The Texas Digital Library (TDL) provides a digital infrastructure for the scholarly activities of Texas universities. The TDL serves as the center of excellence for the creation, curation, and preservation of digital scholarly information including ETDs. The following list outlines the minimum elements for ETDs descriptive metadata for members of TDL. In addition to the following 16 elements (15 mandatory and 1 optional elements), other valid MODS elements may be included in ETD records as appropriate. Detail information regarding the MODS Application Profile for ETD can be found at TDL home page: http://tdl.org.

Title Information
- title
- subtitle

Name of Author
- type=personal
- namePart=given
- namePart=family
- namePart=date
- roleTerm=Author

Name of Thesis Advisor
- type=personal
- namePart=given
- namePart=family
- namePart=date
- roleTerm=Thesis advisor

Name of Committee Member [Optional]
- type=personal
- namePart=given
- namePart=family
- namePart=date
- roleTerm=Committee Member

Name of Degree Grantor
- type=corporate
- namePart=University Name
- namePart=Department
- roleTerm=Degree grantor

Type of Resource
- typeOfResource=text

Genre
- genre=theses
- authority=marcgt

Origin Information
- dateCreated
- dateIssued

Language
- languageTerm

Physical Description
- form=electronic
- internetMediaType=application/pdf
- digitalOrigin=born digital

Abstract

Subject
- authority
- topic
- geographic
- temporal

Identifier
- type

Location
- url=permalink

Degree Information
- name=Doctor of Philosophy
- level=Doctoral
- discipline=Educational Administration

Record Information
- recordContentSource
- recordCreationDate
- recordChangeDate
- recordIdentifier

Bibliography

Agnew, Grace. 2003. "*Metadata Assessment: A Critical Niche within the NSDL Evaluation Strategy*." http://eduimpact.comm.nsdl.org/evalworkshop/_agnew.php.

Alemneh, Daniel. 2008. "Maintaining Quality Metadata: Toward Effective Digital Resource Lifecycle Management." In *Proceedings of the 2008 Knowledge Management: Competencies and Professionalism International Conference*, 313-322, Ohio, USA.

Alemneh, Daniel, and Abebe Rorissa. 2012. "Empowering Digital Libraries Users through Combining Taxonomies with Folksonomies." Paper presented at the Association for Information Science and Technology (ASIS&T) Seventy-Fifth Annual Meeting, Baltimore, Maryland. http://digital.library.unt.edu/ark:/67531/metadc122173/.

Alemneh, Daniel, and Ralph Hartsock. 2014. "Theses and Dissertations from Print to ETD: The Nuances of Preserving and Accessing those in Music." In *Cases on electronic records and resource management implementation in diverse environments*, edited by Janice M. Krueger, 41-60. Hershey, PA: IGI Global.

Alemneh, Daniel, and Samantha Hastings. 2010. "Exploration of adoption of preservation metadata in cultural heritage institutions: Case of PREMIS." In *Proceedings of the American Society for Information Science and Technology*, 47(1): 1–8. http://onlinelibrary.wiley.com/doi/10.1002/meet.14504701187/pdf.

Association of African Universities (AAU). 2014. *Database of African Theses and Dissertations (DATAD)*. http://www.aau.org/?q=datad.

Ballor, Jordan J. 2012. "The Dynamics of Primary Source and Electronic Resource: The Digital Renaissance and the Post-Reformation Digital Library." *ASIS&T Bulletin* 38 (4) (April/May). http://www.asist.org/Bulletin/Apr-12/AprMay12_Ballor.html.

Barton, Jane, Sarah Currier, and Jessie M. N. Hey. 2003. "Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice." In *Proceedings of Dublin Core Conference 2003*, Seattle, Washington. http://www.siderean.com/dc2003/201_paper60.pdf.

Bates, Marcia J. 1998. "Indexing and access for digital libraries and the internet: Human, database, and domain factors." *Journal of the American Society for Information Science* 49: 1185–1205.

Bates, Marcia J. 1989. "The design of browsing and berrypicking techniques for the online search interface." *Online Review*, 13 (5): 407-424.

Beagrie, Neil, and Catherine Pink. 2012. *Benefits from Research Data Management in Universities for Industry and Not-for-Profit Research Partners.* Charles Beagrie Ltd and University of Bath. http://opus.bath.ac.uk/32509/1/RDM_Benefits_vFinal.pdf.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. “The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities.” *Scientific American* 284 (5): 34–43.

Besser, Howard. 2002. "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries.” *First Monday* 7 (6) (June). http://firstmonday.org/issues/issue7_6/besser/index.html.

Bruce, Robert. 2008. “Descriptor and Folksonomy Concurrence in Education Related Scholarly Research.” *Webology* 5 (3). http://www.webology.org/2008/v5n3/a59.html.

Bruce, Thomas R., and Diane I. Hillman. 2004. “The continuum of metadata quality: Defining, expressing, exploiting.” In *Metadata in practice* edited by D. Hillman & E. Westbrooks, 238-256. Chicago, IL: ALA Editions.

Chan, Lois Mai, and Marcia Lei Zeng. 2006. “Metadata interoperability and standardization – a study of methodology, Part I: Achieving interoperability at the schema level.” *D-Lib Magazine* 12 (6). http://www.dlib.org/dlib/june06/chan/06chan.html.

Chevalier, Max, Antonina Dattolo, Gilles Hubert and Emanuela Pitassi. 2011. “Information Retrieval and Folksonomies together for Recommender Systems.” *E-Commerce and Web Technologies* 85: 172-183.

Cleveland, Donald B., and Ana D. Cleveland. 2001. *Introduction to Indexing and Abstracting*. Englewood, Colorado: Libraries Unlimited, Inc. 3rd Edition.

Coyle, Karen and Diane Hillmann. 2007. “Resource description and access (RDA): Cataloging rules for the 20th century.” *D-Lib Magazine*, 13 (1/2). http://www.dlib.org/dlib/january07/coyle/01coyle.html.

Day, Michael. 2006. “The Long-term Preservation of Web Content.” In *Web Archiving* edited by Julien Masanes, 177-199. Berlin Heidelberg: Springer-Verlag.

Day, Michael. 2003. “Preservation metadata initiatives: Practicality, sustainability, and interoperability.” Paper presented at the *ERPANET Training Seminar on Metadata in Digital Preservation.* http://www.ukoln.ac.uk/preservation/publications/erpanet-marburg/day-paper.pdf.

EThOS. 2014. *Electronic Theses Online System: Opening Access to UK Theses*. http://www.ethos.ac.uk/.

Fifarek, Aimee. 2007. “The birth of Catalog 2.0: Innovative interfaces’ encore discovery platform.” *Library Hi-Tech News* (5): 13–15.

Givon, Sharon, and Victor Lavrenko. 2009. “Predicting social-tags for cold start book recommendations.” In *Proceedings of the Third ACM conference on Recommender Systems*. New York, New York.

Global Open Access Portal. 2014. *Cybertesis.* http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/key-organizations/latin-america-and-the-caribbean/cybertesis/.

Golder, Scott A., and Bernardo A. Huberman. 2006. “Usage patterns of collaborative tagging systems.” *Journal of Information Science*, 32 (2): 198-208.

Guy, Marieke, Andy Powell, and Michael Day. 2004. “Improving the Quality of Metadata in Eprint Archives.” *ARIADNE* 38 (January). http://www.ariadne.ac.uk/issue38/guy/.

Hillmann, Diane I., Naomi Dushay, and Jon Phipps. 2004. “Improving metadata quality: augmentation and recombination*,”* In *Proceedings of DC-2004*, Shanghai, China. http://www.slais.ubc.ca/PEOPLE/faculty/tennis-p/dcpapers2004/Paper_21.pdf.

Hughes, Baden. 2004. “Metadata Quality Evaluation: Experience from the Open Language Archives Community.” In *Proceedings, 7th International Conference on Asian Digital Libraries (ICADL2004)*, Shanghai, China. http://www.springerlink.com/content/4kaxeu5p2fb2nac1/fulltext.pdf.

Humphrey, Susanne M. 1989. “A Knowledge-Based Expert System for Computer-Assisted Indexing.” *IEEE Expert: Intelligent Systems and Their Applications* 4 (3): 25-38.

Ivanovic, Lidija, Dragan Ivanovic, and Dusan Surla. 2012. "Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of Novi Sad, Republic of Serbia." *Library Resources & Technical Services* 56 (2): 104-112.

Kim, Hak-Lae, Stefan Decker and John G. Breslin. 2010. “Representing and sharing folksonomies with semantics.” *Journal of Information Science* 36 (1): 57-72.

Lagoze, Carl, and Karin Patzke. 2011. “A research agenda for data curation cyberinfrastructure.” Paper presented at the *11th annual international ACM/IEEE joint conference on Digital libraries – JCDL ’11*, Ottawa, Canada.

Lancaster, E. W. 2003. *Indexing and Abstracting in Theory and Practice*. Champaign, Illinois, Intellitext Corporation.

Lavoie, Brian F. 2004. Preservation Metadata: Challenge, Collaboration, and Consensus. *Microform and Imaging Review 33* (3): 130-134.

Library and Archives Canada (LAC). 2014. *Theses Canada*. http://www.collectionscanada.gc.ca/thesescanada/index-e.html.

Lu, Caimei, Jung-ran Park, and Xiaohua Hu. 2010. “User tags versus expert-assigned subject terms: A comparison of Librarything tags and Library of Congress subject headings.” *Journal of Information Science* 36 (6): 763–779.

Missier, Paolo, Khalid Belhajjame, and James Cheney. 2013. “The W3C PROV family of specifications for modeling provenance metadata.” Paper presented at *EDBT/ICDT* ’13, Genoa, Italy, March 18-22. http://www.edbt.org/Proceedings/2013-Genova/papers/edbt/a80-missier.pdf.

National Information Standards Organization (NISO) 2004. *Understanding Metadata.* www.niso.org/standards/resources/UnderstandingMetadata.pdf.

NDLTDb 2014. *The Networked Digital Library of Theses and Dissertations*. http://www.ndltd.org/.

NISO 2012. *Making Good on the Promise of ERM: A Standards and Best Practices Discussion Paper.* http://www.niso.org/apps/group_public/download.php/7946/Making_Good_on_the_Promise_of_ERM.pdf.

Open Access Theses and Dissertations (OATD) 2014. *Advanced research and scholarship. Free to find, free to use*. http://oatd.org.

OAI-PMH 2014. *Open Archives Initiative – Protocol for Metadata Harvesting.* http://www.openarchives.org/pmh/.

Park, Jung-ran. 2009. “Metadata quality in digital repositories: A survey of the current state of the art.” *Cataloging & Classification Quarterly* 47: 213–228.

Park, Jung-ran, and Yuji Tosaka. 2010. “Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability.” *Information Technology & Libraries* 29 (3): 104–116.

Peterson, Elaine. 2006. “Beneath the metadata: Some philosophical problems with folksonomies.” *D-lib Magazine* 12 (11). http://www.dlib.org/dlib/november06/peterson/11peterson.html.

Phillips, Mark. 2013. Metadata Analysis at the Command Line. *Code4Lib Journal* 19. http://journal.code4lib.org/articles/7818.

PREMIS. 2014. *PREservation Metadata: Implementation Strategies Working Group*. http://www.loc.gov/standards/premis/.

PREMIS. 2012. *PREMIS Implementation Fair 2012: An iPRES 2012 workshop.* http://www.loc.gov/standards/premis/premis-implementation-fair2012.html.

ProQuest. 2014. *DTD to the ProQuest XML feed*. http://www.etdadmin.com/dtds/etd.dtd.

ProQuest. 2014. *ProQuest Dissertations & Theses Database.* http://www.proquest.com/en-US/catalogs/databases/detail/pqdt.shtml.

Sanchez, Elaine R. 2011. *Conversations with Catalogers in the 21st Century*. Santa Barbara, California: Libraries Unlimited.

Saracevic, Tefko. 2006. “Relevance: a review of the literature and a framework for thinking on the notion in information science: Part II.” *Advances in Librarianship* 30: 3–71. doi:10.1016/S0065-2830(06)30001-3.

Schwing, Theda, Sevim McCutcheon and Margaret Beecher Maurer 2012. “Uniqueness Matters: Author-Supplied Keywords and LCSH in the Library Catalog.” *Cataloging & Classification Quarterly* 50(8): 903-928. http://dx.doi.org/10.1080/01639374.2012.703164.

Smith, Gene. 2008. *Tagging: people-powered metadata for the social web*. Berkeley, CA: New Riders.

Spiteri, Louise F. 2007. “The structure and form of folksonomy tags: The road to the public library catalog Information.” *Technology & Libraries* 26: 13-25.

Syn, Sue Yeon, and Michael B. Spring. 2013. “Finding subject terms for classificatory metadata from user-generated social tags.” *Journal of the American Society for Information Science and Technology* 64 (5): 964–980.

Trant, J. 2008. “Studying Social Tagging and Folksonomy: A review and Framework.” *JODI: Journal of Digital Information* 10 (1). http://journals.tdl.org/jodi/article/view/269.

Tunkelang, Daniel. 2009. “Faceted Search.” In Marchionini, G. (2012). *Synthesis lectures on Information Concepts, Retrieval, and services*. DOI: 10.2200/S00190ED1V01Y200904ICR005.

UNT Libraries. “Title Role Authorities.” http://digital2.library.unt.edu/vocabularies/title-qualifiers/.

UNT Libraries. 2014*. UNT Libraries’ Input Guidelines for Descriptive Metadata*. http://www.library.unt.edu/digitalprojects/metadata/descriptive-metadata.

UNT Libraries. 2014. *APIs for UNT Theses and Dissertations.* http://digital.library.unt.edu/explore/collections/UNTETD/api/.

Voorbij, Henk J. 1998. “Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences.” *Journal of Documentation* 54 (4): 466-476.

White, Ryen A., and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. San Rafael, CA: Morgan and Claypool. doi:10.2200/S00174ED1V01Y200901ICR003.

Wichowski, Alexis. 2009. “Survival of the Fittest Tag: Folksonomies, Findability, and the Evolution of Information Organization.” *First Monday* 14 (5). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2447/2175.

Yi, Kwan, and Lois Mai Chan. 2009. “Linking folksonomy to Library of Congress subject headings: an exploratory study.*” Journal of Documentation* 65 (6): 872-900.

Zeng, Marcia L., and Lois Mai Chan. 2006. “Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels.” *D-Lib Magazine* 12 (6). http://www.dlib.org/dlib/june06/zeng/06zeng.html.

Zhang, Ning, Yuan Zhang and Jie Tang. 2009. “A tag recommendation system for folksonomy.” *Proceeding of the 2nd ACM workshop on Social web search and mining*, Hong Kong, China.

7 Guide to ETD Program Planning and Cost Estimation

Gail McMillan (Virginia Tech)

Topics Covered

- Methods to estimate the cost of digital resources.
- Issues with estimating personnel time.
- Issues with estimating technical resources.
- Examples of costs from case studies of ETD programs.

7.1 Introduction

Electronic Theses and Dissertations (ETD) programs rely on a range of administrative and technical resources that require financial support. The objective of this guide is to contribute to an institution's forethought and groundwork and to help stakeholders and decision makers become aware of the potential financial impact of an ETD program. Here we identify costs, but we do not attempt to quantify or calculate them. Institutions need to have an understanding of the costs of the full ETD lifecycle in order to curate them properly and thus ensure their accessibility in both the near and the long term. This guide addresses both born-digital and digitized theses and dissertations, distinguishing between the two only when necessary.

In this guide, planning and cost estimation is associated with the curation lifecycle for ETDs. This lifecycle begins with submission or acquisition and ends with long-term access and preservation of the official document of record. (This document does not consider the costs associated with the creation of ETDs or the digitization of theses and dissertations.) In general, there are two principal financial expenditures: *personnel* and *technology*. The costs associated with personnel fall in to three categories: training, processing, and consulting. The costs associated with technology fall in to three categories: hardware, software, and processing. There can be overlap between these two areas, for example, when processes that are manual could be automated. At this time an *other* category has also been delimited for outlying expenditures such as memberships, fees for outsourced functions and services, or contingency costs.

This document is not intended to be a means to calculate any of these costs. However, interviews with ETD staff at a variety of its higher education institutions resulted in five case studies that share information about their resources and expenditures. The case studies at the end of this document range from new ETD initiatives to mature ETD programs, providing living examples from public and private, large and small, urban and rural universities in the United States.

Since institutional repositories (IRs) have almost become a standard feature of academic libraries, ETDs are just one of an IR's collections. Therefore, this guide does not address "overhead," that is, indirect operating expenses such as facilities (e.g., physical settings for personnel – desks, chairs, electricity, etc., and technology – server room, air conditioning, etc.). Similarly personnel associated with ETD processes usually have other responsibilities as well. Graduate school personnel, for example, who review and approve ETDs, will almost certainly also work on other aspects of graduate student administration such as degree completion evaluations. Similarly, the copyright specialist will provide advice on genre other than ETDs, and the librarian will support the creation of and access to other scholarly works produced and used by the institution's community. This guide does, however, provide estimates for personnel time spent on ETD-related activities, and does so through the case studies.

Though this guide does not prescribe specific personnel positions or technologies necessary for an ETD program, there are a number of major efforts that have studied detailed costs associated with the lifecycle of digital assets with an emphasis on preservation. Many of them provide tools to calculate those costs. There are also several generalized works that would provide a very helpful bibliography beyond the specific references provided with this document, especially Zeller's "Cost of Digital Archiving: Is there an universal model?" (2010). For those seeking tools into which you can plug in expenses, the LIFE3 enables very fine-grained analysis of specific cost components. From 2005-2010 the LIFE Project (Life Cycle Information for E-Literature) created, tested, and made available a Microsoft Excel file and a web tool so that an institution can plug in its expenses in order to determine overall costs. (Hole et al. 2010) In 2005 the National Archives of the Netherlands similarly detailed itemized expenses leading to digital preservation. Another effort, the Keeping Research Data Safe – KRDS, (Beagrie et al. 2010) provides a discounting function when it deploys preservation cost elements. CMDP, the Cost Model for Digital Preservation from the Danish National Library and Archives, also has a somewhat different approach in that it focuses on post facto measuring of preservation costs, rather than predicting costs ("CDMP" 2012). The DataSpace project out of Princeton University proposes a POSF formula – pay once store forever, but covers only the storage costs (Goldstein and Ratliff 2010).

A recent effort to calculate costs for the lifecycle of digital assets comes from the California Digital Library's TCP model – the Total Cost of Preservation. It offers two separate formulas for pay-as-you-go and one-time pricing, but concludes with a hybrid price mode in which the paid-up pricing is used for components that can be quantified and forecast, and pay-as-you-go pricing is used for the others. (Abrams 2012) The TCP model has 10 high-level cost components: (1) services providing curation functions, running on (2) servers deployed by (3) staff in support of content (4) producers who use (5) workflows to submit instances of (6) content types (e.g., ETDs), which occupy (7) storage, and are subject to ongoing (8) monitoring and periodic (9) interventions, all of which are subject to managerial (10) oversight (Abrams et al. 2012). These components relate directly to this document as it outlines the need for cost considerations relevant to personnel and technologies, which will deploy these activities in an institution's ETD program.

This *Guide to ETD Program Planning and Cost Estimation* ascribes to the principles of WORF – Write Once Read Forever (Goldstein and Ratliff 2010), having the same two main objectives: (1) insure that ETDs are and continue to be publicly accessible (if after a very limited restriction or embargo) and (2) cost a reasonable price to store and disseminate. The discussion below will enable an institution to outline its ETD program's need for personnel and technology, and then employ the most appropriate of the tools above or others for cost modeling.

7.2 Personnel

People make policies, design workflows, and provide services. Policies can come from internal institutional decision makers or be derived from peer institutions or aspirational peers. ETD policies are affected by the institution's organization, such as whether there is a centralized graduate school. Policies for ETDs will also be affected by local interest in issues such as Open Access. Unfortunately, policies may also be affected by hearsay, such as myths about publishers' attitudes towards accessible online theses and dissertations, or potential threats such as plagiarism. Direct costs are rarely associated with policy making and are considered as part of the general expense of operating a program.

Workflow can be designed internally, or it can be adapted from existing ETD programs at other institutions. It may also be predicated upon vendors' software and hardware requirements. Costs can be associated with the people designing the workflow or with the system chosen to implement the ETD initiative. Sources for adapting workflows include members of the Networked Digital Library of Theses and Dissertations or the census of required ETDs.[1] Workflow can be prescribed by:

- External commercial enterprises such as ProQuest and Bepress.
- Open source digital repository applications such as Vireo for DSpace[2] and OpenETD from RUCore.[3]
- Consortial memberships such as OhioLink[4] and Florida Virtual Campus.[5]

Maintaining an ETD program will rely heavily on the systems administrator, but the personnel in these positions are usually assigned several systems to oversee. According to the case studies that follow, whether the hardware is outsourced or locally maintained, the time commitment is small though constant. Similarly, whether an institution employs open source or commercial software, even the most out-of-the-box software will require some modification to meet an institution's specifications, workflow, quality control and the like.

Personnel are heavily involved in training during new ETD initiatives. Even when face-to-face instruction is replaced by online, asynchronous tutorials or self-explanatory submission and processing instructions, personnel will be needed to work with individuals (e.g., authors, users, staff, and faculty) as issues come to light. Mature ETD programs may go for years without needing to modify instructional documentation but the migration from legacy to modern systems usually requires new instruction and rewritten guides.

[1] See NDLTD, http://www.ndltd.org/, and ("Institutions Requiring Electronic Thesis/Dissertation Submissions", https://docs.google.com/spreadsheet/ccc?key=0AtSglIhGWCkpdHJvOUNSZUZyRC04UXRUa0w3UmgtYWc&hl=en#gid=0.

[2] See http://sourceforge.net/projects/vireo/.

[3] See http://rucore.libraries.rutgers.edu/open/projects/openetd/.

[4] See http://etd.ohiolink.edu/.

[5] See https://fclaweb.fcla.edu/node/528.

Institutions vary as to whether these activities are the responsibility of personnel in the graduate school or the library.

The most time-consuming personnel activity may be reviewing and approving each individual ETD. Though the author's committee approves the plan of study, which leads to their sanctioning the final document, graduate school staff (almost always) also review ETDs for conformity to their institution's requirements and guidelines. Since ETDs are still almost entirely text-based documents, most institutions continue to prescribe page layout as they did when they needed to conform to bindery specifications. Page number placement, margin widths, and the like continue to be a tedious part of the ETD review process.

The case studies that follow document that reviewing and approving an ETD can take as little as 15 minutes, or it can stretch to more than 30 minutes when the reviewer must email the author about needed changes and re-review the subsequent versions. Some institutions assess fees on ETD authors based on the amount of time the graduate personnel spend reviewing the work. For example, University of South Florida's ETD FAQ states, "graduate school fees are applied based on the cost of review. Submissions that require extensive editing or repeated review *will* accrue additional processing costs. (USF 2013)

While students are usually allowed to submit their ETDs at any time, the graduate school's deadlines for graduation[6] generate peak periods of ETD submissions. Most ETDs are submitted online, but some graduate school personnel are responsible for receiving these works on portable media (e.g., CDs, flash drives, etc.). Even in the "off season" the back-and-forth emails between staff and authors can be time-consuming. The larger issue of quality control may require manual review of individual documents, but technology is beginning to be used to some extent to ensure that ETDs are in approved file formats (e.g., PDF/A) (see also *Managing the Lifecycle of ETDs: Curatorial Decisions and Practices*).

There are personnel costs associated with copyright monitoring and compliance, and may include intellectual property rights consulting. Institutions must consider the costs associated with determining legal rights to include copyrighted works in an ETD, certifying an author's original work, and managing the institution's non-exclusive license to store and provide access to ETDs (see also *Briefing on Copyright and Fair Use Issues in ETDs*).

Personnel with liaison responsibilities to students provide face-to-face, phone, and online consulting as well as classroom and online instruction. Some graduate schools offer walk-up services during regular business hours so that authors can bring their documents on portable media for an informal review and preliminary approval. However, personnel have largely been removed from the final approval notification process. Programmed scripts can automatically notify authors and their committees through emails when the graduate school has approved their ETDs.

[6] e.g. Virginia Tech's, http://graduateschool.vt.edu/academics/dates_deadlines/commencement_deadlines.html.

Another largely avoidable personnel cost is in the movement of files from restricted and embargoed status to open access. Even when this can be done programmatically based on the calendar, there are circumstances that require individual attention. The library may receive inquiries from patrons who are not members of the university community and want to have access to restricted ETDs. Sometimes students have heard about an ETD but can't find it because the author chose to have it embargoed. These circumstances often result in the library's ETD liaison individually locating email addresses and contacting authors of the restricted and embargoed ETDs for permission to remove the restrictions and provide public access prior to the timed release. When the liaisons receive permission (as they usually do), they must login to the system and change the status of the ETD. This is often an opportunity to improve the metadata or its display. Occasionally, at the request of the appropriate authority (e.g., dean of the graduate school), it is necessary to move an ETD temporarily from open access to restricted or embargoed status (e.g., data errors discovered after approval, plagiarism claims, etc.).

Another personnel cost is cataloging ETDs and linking URLs to the bibliographic records in the library's online catalog. Many institutions have avoided this cost by programmatically deriving the MARC (machine readable cataloging) record from the ETD's metadata. Similarly, some institutions derive metadata from the MARC record to describe the scanned theses and dissertations in their institutional repositories. However, quality control and authority control are often a librarian's responsibility, requiring manual look-up and online editing.

Most American universities partner with ProQuest/UMI to ensure that their ETDs are available through its various commercial products, even when these institutions also make their ETDs available in open access repositories. ProQuest accepts ETD submissions through its online portal, the UMI ETD Administrator. There is no charge for using the ETD Administrator, but universities submitting theses and dissertations on paper and other media or uploading PDFs through FTP incur a $25 charge per ETD. When submitting permission forms and PDFs outside the ETD Administrator, ProQuest requires an original signature, leaving university personnel to collect and mail these forms. Though payments are increasingly made online, universities continue to manually audit ProQuest invoices.

Over time many personnel responsibilities will be automated by technology. Customizing software is one example where we are able to interface our human resource systems such as Banner and PeopleSoft with graduate school approval systems and institutional repository logins. Like many of the issues mentioned above, these are not necessarily particular to ETDs but can apply to other genres and IR collections.

7.3 Technology

Technology costs are easier to estimate for hardware and software. But personnel run the hardware, customize and/or maintain the software, and oversee the systems and processes.

In planning and considering the costs associated with the lifecycle of ETDs, there should be a continuum of curation, a "consistent and coherent regime of management processes from the time of the creation through to preservation and use," (Netherlands (2005), AS 4390.1). It is preferable to assign as much to technology and as little to personnel so that staffing changes have less effect and processing is timely and quick. As stated earlier, we are not including overhead expenditures such as buildings and rooms, electricity, air conditioning, heating, and the like, because these are largely in place and already host the people and technology necessary for an ETD program. Even the largest ETD programs find that they are a small addition to an existing IR or digital library.

Whether ETDs are going to be a stand-alone program or added to the institutional repository with other digital works, we are providing planning and cost estimates for necessary technologies. An early decision needs to be whether to host ETDs locally and/or contribute them to an external entity such as a consortium (e.g., OhioLink) or a commercial service provider (e.g., ProQuest), and whether to store the institution's ETDs locally (i.e., on campus equipment) or remotely (e.g., cloud storage, service provider), or to employ a combination of these options. This document addresses a variety of options as broad technology topics. (see also the chapter *Guide to Options for ETD Programs*)

7.3.1 Local Hosting

When ETDs are a new genre of digital works to be processed and managed, technological considerations include hardware and software that may be hosted by the institution's library or by its central information technology (IT) unit. Technology can be acquired to handle submission, storage (including maintenance and back-ups), access, and preservation. (See above for the human part of this process, which will minimally cover the ETD approval process. Desktop computers with adequate capacity for timely processing are essential.) Network technology is also necessary for the communications among the hardware, software, and personnel.

The hardware required will include servers for processing (submission and/or intake; approval and/or evaluation), storage (whether temporary (i.e., just through approval) or long-term), backups, access, and preservation. Recommended storage media are spinning discs, but some may use them in combination with tapes. However, we do not recommend compact disk storage.

To reduce risk and increase productivity, multiple servers will serve the institution well for purposes of development and testing as well as processing, access, and preservation. Network communications should be robust (high speed connections) and flexible (bandwidth) enough to adequately handle timely transfer of files of increasing size. As ETD programs mature, many of these works will evolve beyond text-based documents to include more or even become multimedia works. Along with this evolution comes the need for institutions to not only store multimedia but also to stream it. We are in the midst of the shift from users not having a choice about downloading files for local access onto their computers to users opting for access to streaming video and audio files that remain on institutional servers.

Other issues that will affect the server size and capacity are the challenges of migration and emulation, whether at ingest or at some other point in the life cycle have yet to be played out. Some institutions will also want servers and storage equipment that will enable them to be part of research and development endeavors or just for testing. For example, an institution may anticipate hosting large datasets that could accompany ETD submissions, and they must have the capacity to sufficiently test the success of various analysis, manipulation, and viewing tools.

An institution's technology may be in place if it is already hosting digital works. Adding ETDs could be just another genre of works to be made available through the digital library or institutional repository. ETD submission could be a variation of the faculty deposit process, with student works enjoying a more formal or at least different vetting process.

The NDLTD's *ETD Guide* (NDLTD, 2011) recommends a common server concept for an ETD program. It includes four servers, one each for production, back up, public access, and preservation. The production server (1) communicates with the users, authors, graduate school personnel, etc., and the uploading and metadata processing is done here. The back-up server (2) is sometimes called the archive server. It ensures that a continual incremental backup of the public archive server. Only system administrators have access to this server, securing access to the ETDs and their digital signatures (i.e., their authenticity and integrity). The document server (3) is also known as the public archive server because users access and retrieve ETDs here. Secure this server with a RAID (Redundant Array of Independent Disks) system that internally holds two equal copies of the server distributed on independent hard disks to ensure that if one copy cannot be accessed due to a head crash or other hardware inconsistencies, the second copy will take over functionality and operate as the first copy. Users will not realize which copy is in use. Lastly, the preservation server (4) provides another storage management system--an archive that hosts the original ETD files, their metadata, and management software. It is important if an environmental accident such as an earthquake or human error destroys the original archive, that a preservation server be located at another geographic location, though ideally several institutions will collaborate to host preservation servers for each other. (see also the chapter *Guide to Options for ETD Programs*)

7.3.2 External Hosting

It may be an option to shift much of the technical capacity suggested above to an external host or agent. For example, if the institution is part of a statewide consortium there may be little need for developing a local infrastructure because the consortium may not only store, but also process ETDs, and provide access as well as preservation services.[7]

Integrated library services that handle ETD processing and hosting are also commercially available through a variety of vendors such as:

- Bepress Digital Commons, used at University of Massachusetts at Amherst, Pace University, and University of Iowa.[8]
- Ex Libris DigiTool used at Florida State University and Universidade do Porto (Portugal).[9]

These systems may process digital works in addition to ETDs.

Commercial ETD processing and hosting services are also available. ProQuest/UMI is the most well known one in the US, with its ETD Administrator.[10] It also offers to facilitate local access in combination

[7] Examples include OhioLink (http://etd.ohiolink.edu/) and the Florida Library Virtual Campus (http://fclaweb.fcla.edu/node/528).

[8] See http://scholarworks.umass.edu/, http://digitalcommons.pace.edu/dissertations/, and http://ir.uiowa.edu/.

[9] See http://www.lib.fsu.edu/find/etds.html and http://www.exlibrisgroup.com/category/DigiToolOverview.

with remote processing by returning an institution's ETDs for local hosting. See the University of Arizona case study below.

7.3.3 Software

Of course the equipment requires software that will operate the system, enable it to interact with other systems, and process individual ETDs either minimally or extensively. A consideration in purchasing equipment is the operating system that will run on it and its compatibility with other systems.

ETD-specific software is available from a variety of sources. Open source options include:

- *ETD-db* from Virginia Tech.[11]
- *Valet* from VTLS.[12]
- *Vireo* for DSpace from the Texas Digital Library.[13]
- *OpenETD* from Rutgers.[14]
- *Cyberteses*, which is popular in French and Spanish speaking countries.[15]

These were summarized in a 2011 article by the newest entry in the ETD software market, Jarrow from the University of Northern British Columbia. (MacDonald and Yule, Daniel 2012) In addition to these, the 2008 ETD preservation survey[16] revealed that other platforms and repository structures that also support ETD programs. According to the ETD preservation survey, other software in use includes:

- EPrints at Cal Tech.[17]
- CONTENTdm at Brigham Young University.[18]

Library system vendors are continually joining the market place. For example, BePress offers an ETD toolkit with its Digital Commons repository software.

7.3.4 Other

Costs associated with ETDs that do not fall into the personnel or technology categories include the association memberships, marketing, and contingencies. Institutions should consider membership in organizations such as the Networked Digital Library of Theses and Dissertations (NDLTD), national groups such as the USETDA – United States ETD Association, and statewide consortia (e.g., Texas ETD Association). Membership fees are quite low and provide many opportunities to learn from and

[10] See http://www.etdadmin.com/cgi-bin/main/about.
[11] See http://scholar.lib.vt.edu/ETD-db/index.shtml.
[12] See http://www.vtls.com/products/valet and a live instance in place at the University of Hong Kong http://etd.lib.hku.hk/.
[13] See http://sourceforge.net/projects/vireo/files/.
[14] See http://rucore.libraries.rutgers.edu/open/projects/openetd.
[15] See http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/key-organizations/latin-america-and-the-caribbean/cybertesis/.
[16] See http://lumiere.lib.vt.edu/surveys/results/view_results.php3?set_ID=ETDpreservation200708.
[17] See http://thesis.library.caltech.edu/.
[18] See http://contentdm.lib.byu.edu/cdm/search/collection/ETD.

contribute to ETD initiatives at peer institutions and offer discounts at national and international venues. Personnel costs might include attending the conferences of these associations.[19]

Another cost is marketing, for example, a new initiative, a successful program, award winning ETDs, and the like. Contingency costs should also be considered. These would include unanticipated expenses associated with trigger events, catastrophes, and sea changes, and will vary depending on the size of the ETD program and its complexity.

7.4 Case Studies

This document has highlighted anticipated expenses associated with new and mature ETD programs. It has generalized personnel and technology needs in order to inform stakeholders and decision makers about potential resources requirements. In this section we offer five case studies from institutions large and small, private and public, rural and urban with either new ETD initiatives or mature ETD programs. They are the result of interviews with representatives at Portland State University, Rice University, University of Arizona, University of North Texas, and Virginia Tech in October-November 2012. Each interviewee responded to several broad questions meant to identify a general set of cost factors and those that are most critical to ETD programs at various stages of planning. See Appendix A for the individual responses from each institution. The case studies shed light on the financial information stakeholders in a new ETD initiative or an existing ETD program would fine relevant, and they revealed what is common across institutions of higher education and what issues may be unique.

The five case studies provide examples of how these universities approached their ETD initiatives and specifically identify their technology, personnel, and resource allocation issues. Though the universities vary, their ETD programs have many similarities. The library hosts ETDs at four out of the five universities, while one is hosted by central IT (Portland). Each university has three-to-four graduate school staff who review ETDs. None of the universities provided supplemental resources to the library or the graduate school for their ETD programs and each has absorbed ETDs into the routine activities of the library and the graduate school.

There are two fairly new ETD programs at Rice and Portland, which required online submission in 2011 and 2012, respectively. Virginia Tech and the University of North Texas have long-standing programs begun in 1997 and 1999, respectively, while Arizona's began in 2005. It may be noteworthy that the two oldest ETD programs (VT, UNT) have locally developed ETD systems, while the newer programs use DSpace (Arizona, Portland, Rice). VT is in the midst of migrating from its ETD-db system to DSpace, and is already planning for a Fedora platform. The ETD programs at Arizona and Rice employ remote services, choosing to pay for cloud storage or a vendor-hosted institutional repository rather than local equipment and staff.

[19] See http://txetda.wordpress.com/etd-forum/2012-txetda-annual-conference/ for the Texas ETD meeting in 2012, annual US ETDA conferences: http://www.usetda.org/?page_id=1591; and NDLTD sponsored international conferences at http://scholar.lib.vt.edu/theses/ndltd.html.

Graduate students submit their ETDs through the ProQuest ETD Administrator at Arizona and Portland so their students are not charged for having their ETDs included in the ProQuest Dissertations & Theses Database. Rice, UNT and VT's students are charged a $25 ProQuest fee to have their dissertations included in the ProQuest database. Two universities also charge their graduate students an archiving fee of $20 (North Texas, VT), while Arizona charges a $6 ETD fee. But Rice, a private university, and Portland, a public university, do not charge graduate students additional fees. They also have the smallest number of ETDs approved annually--300 at Rice and 200 at Portland. UNT receives about 400, Arizona 600-700, and Virginia Tech receives 800-900 ETDs annually. Digitizing bound theses and dissertations that precede the ETD requirement is another common activity disclosed in three of the five case studies. Rice is an exception, purchasing its dissertation files from ProQuest.

While the case studies reveal many commonalities across institutions, the case studies also highlight specific local resources. The Virginia Tech case study is unique in that it details the key technologies used in their ETD program and the costs. Arizona's case study reveals its remote server costs and details per-title expenses. At Portland the Office of Graduate Studies plays a larger role in the ETD program than the other universities, which reveals the extensive role of the libraries in most ETD programs. The workflow in the Rice case study stands outs because it includes cloud storage. While ETD workflow is generally the same at each institution, it is noteworthy in the UNT case study that the Graduate Reader receives ETDs delivered by graduate students on portable media (e.g., CDs, flash drives). After she reviews and approves them, she compiles them on a CD for the library where the PDFs are converted to high resolution JPGs and optical character recognition applied (OCR). This provides an added level of quality control as well as a page-view interface that is not part of processing at the other universities.

7.5 Summary

Institutions need to understand the costs of the full ETD lifecycle in order to curate them properly and thus ensure their accessibility in both the near and the long term. There are two principal financial expenditures: personnel and technology. The costs associated with personnel fall in to three categories: training, processing, and consulting. The costs associated with technology also fall in to three categories: hardware, software, and processing. Though this guide does not prescribe specific personnel positions or definitive technologies necessary for an ETD program, there are a number of major efforts that have studied detailed costs associated with the lifecycle of digital assets that may influence ETD programs, namely: the LIFE Project (LIFE3), Keeping Research Data Safe (KRDS), Cost Model for Digital Preservation (CMDP), DataSpace Project, and California Digital Library's Total Cost of Preservation (TCP). In addition to covering the broad landscape of costs and resources for coming to terms with ETDs, this guide also offers five case studies that provide examples of how a range of American universities have approached their ETD initiatives and specifically how they have identified technology, personnel, and resource allocation issues.

Appendix A: Case Study Interviews

1. Portland State University Case Study Interview

Courtney Ann Hanson, Coordinator of Graduate Studies

Does your institution have an ETD program or does it have a planning group considering an ETD initiative?

Portland State University began with limited testing in Summer 2009, gradually expanded the program, and transitioned to ETD-only submissions in Summer 2011. No paper theses and dissertations are accepted.

What are the key technologies considered necessary to initiate and/or maintain, develop, and preserve an ETD program?

- Submission tool. We use the ProQuest ETD Administrator, which provides a four to six week turn around.
- File formats. Accepts all file formats, though largely PDFs Received audio and Excel files as supplemental materials.
- IR. We currently use DSpace, but will probably soon transition to Digital Commons from Bepress.
- Servers for backup and preservation are the responsibility of the university's central IT department.
- ProQuest ETD Administrator. Portland uses this system in its entirety. Students largely select from the PQ access levels and the IR programmatically removes restrictions based on the OGS approved date. The only ETDs that are university-only accessible are from the Creative Writing department. Portland doesn't allow embargoed ETDs, but would consider extending restricted (i.e., university-only) access to five years if necessary.

Who are the primary university personnel involved in or considered rudimentary to your ETD program and in what units do they work?

In the Office of Graduate Studies (OGS):

- Coordinator of Graduate Studies is the point person in OGS for all things ETD. She is a liaison to students, faculty, and the library. After devoting considerable time to launching the ETD program and establishing policies and procedures, they now take less than 10% of her time. [There have been very few instances of changing the availability of an ETD outside of the programmed releases.] She does handle paper forms for approvals, access levels, and ProQuest permissions.

- Three degree processors review and approve ETDs as part of the degree audit process. This may take 5% of their time. While 1900 students complete their graduate degrees at Portland in a typical year, only 200 submit ETDs.

In the library:

- Digital Initiatives Coordinator [Karen Bjork] is new to this position (circa one year) and is the counterpart of the Coordinator of Graduate Studies.
- Scholarly Communications Librarian helps with workshops, copyright, open access, etc.
- Metadata Cataloger supervised by the Head of Cataloging. The library catalog links to the ETD in the institutional repository.

What other expenditures are associated with your institution's ETD program?

- ETD conference attendance
- US ETD Association membership fee

How are resources allocated to your institution's ETD program? In the past three years have your costs in the categories of technology, people, and other: Increased? Decreased? Stayed roughly the same? Why?

Resource allocations for ETDs at Portland have stayed the same, which is they continue to operate with the on-hand resources – staff, systems, etc., as they have from the launch of the program. The work has been incorporated into existing positions and no new positions were created to support ETDs, either in OGS or the library. Portland selected the ProQuest ETD Administrator as its submission tool not only because the product was free, but IT support was also provided by the university's central unit at no cost.

Portland is currently using the open source software DSpace. They are likely to switch to Digital Commons soon but this is not related to their ETD program.

How are costs shaping the current planning and direction for your ETD program?

See above.

2. Rice University Case Study Interview

Geneva Henry, Executive Director, Digital Scholarship Services (DSS)

Does your institution have an ETD program or does it have a planning group considering an ETD initiative?

Yes. In 2006 the Rice library purchased its dissertation files from ProQuest and made them available locally through its IR. RU was still only accepting print theses and dissertations. Henry worked with Graduate Studies to receive files from students who wanted open access. She was the contact on the Graduate Studies web site, and she provided online submission instructions for students, but this was in addition to the Graduate Studies requirements.

During the 2011/12 school year, ETD submission became an official option. DSS had tested Vireo and her programmer had customized it to meet Rice's needs. They received some online submissions fall semester, but spring semester the number increased considerable because it was both easier and less expensive to submit online. ETDs became mandatory with 2012/13 school year. The library receives approximately 300 ETDs annually.

Access levels: ETDs with embargos are flagged when they are imported into Vireo and automatically released when their due dates arrive. Rice does not manually release ETDs early and they have not yet had requests for delaying releases, though they would honor these requests for relevant reasons.

What are the key technologies considered necessary to initiate and/or maintain, develop, and preserve an ETD program?

Open source software, Vireo, was customized heavily, as every institution must do to incorporate individual workflows. For example, Graduate Studies has forms, such as the certificate of completion, that are routed to the institutional records management system (handled by the university administration, not the library), while the ETD is sent to the library's institutional repository (using SWORD, (Simple Web-service Offering Repository Deposit).

The Rice repository is DSpace, but looking to the future may be including a Fedora foundation such as Hydra or Islandora. ETDs are very little trouble – require small amount of storage space, automated processes run smoothly with current policies, etc.

Rice hosts the hardware for the repository, but chooses to run Vireo for ETDs in the Amazon Cloud. The monthly cost for cloud storage is relatively low considering the server and maintenance costs that would be incurred if it were managed locally. The other significant cost with cloud storage is with delivery of content out of the cloud. There are really just two "bursts" of activity each year: when the students are submitting their theses close to the deadline each semester. It is very cost effective because they receive files from the Graduate School only once a year after May graduation, which replicates their previous practice of shipping theses and dissertations to ProQuest annually. Every May when students graduate, ETDs are dropped into the Rice repository.

The library has in-house servers for testing and development (e.g., Vireo customization), but the production servers, like the IR, are run on servers in the primary data center for the university.

Who are the primary university personnel involved in or considered rudimentary to your ETD program and in what units do they work?

Graduate Studies has three ETD Reviewers, who process about 300 ETDs per year.

Henry, a programmer and a digital curation coordinator, is the principal staff involved with ETDs in the library. Her unit, Digital Scholarship Services provides support for ETDs, such as the research and testing, which led to using Vireo for ETDs. Library staff from the Acquisitions Department handles ProQuest billing. Technical Services could become involved by handling metadata quality control. The IR could export to MARC if necessary. Henry recommends these four types of staff be involved in an ETD program.

What other expenditures are associated with your institution's ETD program?

None beyond what was mentioned above: purchasing files from ProQuest, Amazon cloud, and local personnel and hardware.

How are resources (going to be) allocated to your institution's ETD program?

From the library point of view, ETDs are just a normal part of the workflow. Library has automated all aspects of ETDs, and nobody has to be assigned to specifically look after them. When they were developing the online workflow and adapting Vireo, a programmer had primary responsibility and it was a priority but not the only work assignment. Vireo development and testing took three months.

The Graduate Studies office has the same basic procedures but now they do it online.

In the past three years have your costs in the categories of technology, people, and other: Increased? Decreased? Stayed roughly the same? And why?

Costs increased with Amazon Cloud, but no additional personnel, hardware, or software costs.

How are costs shaping the current planning and directions for your ETD program?

A major concern is to reduce costs for students, but they must still pay the $25 ProQuest fee largely because Graduate Studies considers ProQuest to be the best source of information about dissertations in the United States. There is concern, however, that students are not reading the ProQuest forms. [It was also a concern when ProQuest contracted with third party vendors to market ETDs, but ProQuest discontinued these commercial relationships. (per Austin McLean, July 9, 2013)]

3. University of Arizona Case Study Interview

Yan Han, University of Arizona Associate Librarian for Research Services

Does your institution have an ETD program or does it have a planning group considering an ETD initiative?

Yes. University of Arizona started to discuss implementation of an ETD program in 2005. It started with a conversation between the Graduate College and the library for e-submission.

Yan Han, who was on the digital library team, led an informal planning group that included library staff from cataloging, technical services, digital library, and document delivery. He drafted a document about the impact of e-submission and access, estimating costs and fees, and the dean and department heads provided their input. Subsequently, the library and the graduate college agreed to begin accepting ETDs in October 2005 using the ProQuest e-submission platform.

The Graduate College is responsible for managing e-submission, while the library takes care of the rest of the ETD life cycle. The Graduate Collect negotiated with ProQuest to receive copies of the ETD files for the Arizona library's server. The Graduate College collects the signed distribution rights release forms from the graduate students, which give the library permission to make their ETDs open access. This is the only paper in this process.

Han wrote scripts to convert the ProQuest metadata into the MARC bibliographic record for the library catalog, and also to prepare the ETD files for ingesting into DSpace.

What are the key technologies considered necessary to initiate and/or maintain, develop, and preserve your ETD program?

Arizona uses ProQuest's ETD Administrator; therefore, the only equipment the Arizona library must maintain is an FTP server to receive files from PQ. Arizona pays $24,000/year for a remotely hosted DSpace server (one-third the costs of a systems administrator). The only software to maintain is what Han wrote to convert xml to make MARC load into the library catalog and repository (DSpace).

Who are the primary university personnel involved in or considered rudimentary to your ETD program and in what units do they work?

Graduate College representatives include the Associate Dean and the Supervisor of Certification who oversees four staff. These were the same staff positions that handled paper theses and dissertations. They review student submissions for Theses & Dissertation manual such as margins, page numbers, etc.; approve; receive signed copyright permission forms. [This is the same situation as at Virginia Tech.]

Library working group had representatives from cataloging, technical services, digital library, document delivery, Han, and later dean and department heads. Library has a copyright librarian to handles those questions from ETD authors, and a technical person (Han) to help students create PDFs. Library staff are not involved in submission process but receive paper copyright forms to digitize. Cataloging is done programmatically. Two library staff with student assistance handle access and preservation. Han manages loading PQ files; systems administrator preserves them.

Loading takes about 10 minutes per ETD. Arizona processes 600-700 ETDs/year.

Access levels per PQ are: no restriction, 6 months, 1 year, 2 year, or a specific future date.

- 98% of Arizona's ETDs are open access from Arizona repository.
- Restrictions are programmatically removed.
- 5-10 cases/year Han manually changes the restrictions.

What other expenditures are associated with your institution's ETD program?

Two library colleagues handle TD back file digitization.

Processing PQ invoices includes searching the library catalog but as of 2011 PQ no longer charges since Arizona uses the PQ ETD Administrator.

When students use PQ for copyright registration, library still processes those invoices.

Hosted DSpace server is $24,000/year.

How are resources (going to be) allocated to your institution's ETD program?

Graduate College processing hasn't changed much from paper TDs to ETDs, but save a little by not handling paper.

Library experiences the major cost savings.

2005: student's dissertation fee was $135: $85 ProQuest fee for dissertation listing, $35 candidate fee, and $15 dissertation processing. Han estimated library cost of $7 per ETD for staff and supplies.

2012: $13,000/year to host ETDs (born-digital and digitized TDs). ETDs are 55% (10,600 titles, half are digitized TDs) of the IR collection (T=19,000, $24,000/year for hosting). Therefore, $1.80/title/year. Graduate College staff spend 30 minutes reviewing the average ETD. Han spends 30 minutes a month.

Arizona graduate student ETD fee=$3: hosting at $2/title/year; preservation at $1/title/year. The library supports these costs for all subsequent years.

In the past three years have your costs in the categories of technology, people, and other:

Increased? Decreased? Stayed roughly the same? And why?

Decreased due to ProQuest eliminating fees to libraries that use its ETD Administrator.

How are costs shaping the current planning and directions for your ETD program?

Processing is going smoothly and costs are minimal, so no need to consider changes.

4. University of North Texas Case Study Interview

Daniel Alemneh, Digital Curator, Libraries

Mark Phillips, Assistant Dean for Digital Libraries

Jill Kleister, Graduate Reader, Toulouse graduate school

Does your institution have an ETD program or does it have a planning group considering an ETD initiative?

Yes. The UNT program started in 1998, but ETDs were not required until 1999.

What are the key technologies considered necessary to initiate and/or maintain, develop, and preserve an ETD program?

Until 2007 the graduate school and the university's central IT unit handled ETDs. From 1999 to 2007 the library was making lots of changes to its technological infrastructure. In 2007 the library took over ETDs from central IT and began working directly with graduate school. ETDs became part of the locally designed workflow for digital library resources.

UNT designed its digital library systems using open source software to meet its identified needs. Because ETDs are part of UNT's digital library resources, hardware and software is not specific to ETDs but to the needs of the UNT digital library. Seed money for the UNT digital library came from a variety of internal and external resources, including federal grants, state subsidies, and university awards.

When the library receives the ETDs, they are PDF files. The library converts these to high resolution JPGs, OCRs the JPGs, and creates a page-view interface. ETDs are available in the UNT digital library as both the original PDFs and the JPGs. Occasionally an ETD is accompanied by auxiliary files and these are packaged with the ETD.

Who are the primary university personnel involved in or considered rudimentary to your ETD program and in what units do they work?

Daniel Alemneh oversees ETDs, among other duties. When there are no special ETD-related projects (unlike this year with the IMLS grant, theses and dissertations digitization project, and locating pre-2007 ETD authors), he would spend at the most 20% of his time on ETDs. He is the liaison with the graduate school and oversees the processing of about 400 ETDs annually: http://digital.library.unt.edu/explore/collections/UNTETD/browse/. The library staff does not work directly with students or faculty on ETDs. However, the library recently hired a scholarly communications librarian who will become involved with the university community, including ETD authors.

Jill Kleister is the "Graduate Reader" and the only staff in the graduate school who interacts with graduate students and faculty, reviewing and approving ETDs. Graduate students deliver their signed (1) ETD approval forms, (2) ProQuest permission forms and sometimes copyright registration forms, and (3) ETDs on CDs and flash drives. She uploads the ETDs to her computer, which are almost always PDF files, then reviews and approves them. Each semester she compiles a CD of approved ETDs and sends the CD to the library. After the library has performed some quality control and uploaded the semester's ETDs to the institutional repository, they compile a final CD and send it to Jill. Jill prints each title page and abstract and matches these with the ProQuest form(s) and sends them with the library CD of ETDs to ProQuest. The graduate school recently added .5 FTE to assist Jill with her job duties.

What other expenditures are associated with your institution's ETD program?

Digitizing bound theses and dissertations (BTDs), which is largely being done by two students and overseen by Alemneh. It is expected to be a three year project to digitize about 10,000 BTDs. Scanned TDs will be open access, closed if author requests. Texas Tech did this without any problems.

Locating circa 700 ETD authors to request permission to make restricted (or indefinitely embargoed) ETDs open access. Will use Qualtrix as a click through agreement. Graduate school is working on locating every author's email address. Students will be asked to complete a unique survey, which will be a request for permission to make the publicly accessible. Alemneh oversees this project:

http://digital.library.unt.edu/explore/collections/UNTETD/browse/?fq=dc_rights_access%3Aunt.

Automating the ETD submission process to replace UNT graduate students delivering CDs to the graduate school. Alemneh will work with the graduate school's programmer.

How are resources (going to be) allocated to your institution's ETD program?

The library supports ETDs after the graduate school has approved them. The library performs additional processing to provide access to the original files but also to derivatives (data desiccation, e.g., PDF to image files). [http://digital.library.unt.edu/ark:/67531/metadc67625/] PDF files are converted into high resolution JPGs; JPGs are OCR'd, creates a page view interface. Store pages and PDFs, if auxiliary files add audio, etc. files with the ETD package. PDF and metadata file are converted into local metadata format, manual (?) conversion from xml to html, infrastructure helps with metadata construction. Public or restricted ETDs are in one collection, with single search access and various views, e.g., digital library or portal.

The graduate school handles receiving, reviewing, and approving ETDs, sending the ETD files to the library three times per year, and sending the ETD files and the signed forms to ProQuest. The graduate school has assigned a programmer to automate the submission process. (GS (and library?) is looking at open source software such as Vireo and Drupal, but UNT has Drupal expertise.)

In the past three years have your costs in the categories of technology, people, and other: Increased? Decreased? Stayed roughly the same? And why?

Costs have risen but it is a temporary situation until the projects listed in "other expenditures" are completed.

UNT wants to increase graduate student enrollment but this will not have a big effect on the library's ETD program as the automated processing is in place and functioning well with little oversight or need for problem solving.

The library is eager to support all file formats but the graduate school requires a single PDF file ["Legacy of limits"]. With a more automated system they could readily accept audio, video, CAD files, etc. Students are beginning to inquire about including other content with their ETDs. These files are now occasionally accepted at the discretion of the graduate school Reader working with individual students. PDF with imbedded audio and video is a challenge.

How are costs shaping the current planning and directions for your ETD program?

Library has assumed the costs associated with scanning BTDs and locating authors to remove restricted (embargoed?) access.

The graduate school will assume the costs for programming to automate ETD submission.

Students pay two fees as part of the graduation application: $25 "microfilm" fee and $20 "archive" fee. The $20 microfilm fee is what ProQuest charges the UNT library. The archive fee also goes to the UNT library to cover costs associated with the long-term maintenance of ETDs. UNT is currently investigating whether to continue mandatory ProQuest deposit.

5. Virginia Tech Case Study Interview

Gail McMillan, Director, Digital Library and Archives (DLA)

Paul Mather, Systems Administrator, [Library] Information Technology and Services

Kimberli Weeks, Technical Director, Digital Library and Archives

Janice Austin, Director of Graduate Admissions and Academic Progress

Does your institution have an ETD program or does it have a planning group considering an ETD initiative?

Yes. Virginia Tech as the first university to require ETDs (Jan. 1, 1997). We began drafting online workflow in 1995 and had a pilot in place in early 1996.

What are the key technologies considered necessary to initiate and/or maintain, develop, and preserve an ETD program?

2 x 500 GB drives (as RAID 1), in a four-drive chassis, to allow for easy storage expansion. The current system, serving our ETDs is a dual quad-core system with 8 GB of RAM, so we went with the same number of CPUs and RAM. [Prices are mainly from GovConnection.com, which many higher-education institutions use. The server quotes are from eRacks and Dell, which give a range of prices from likely vendors.]

$2500--$2700 Server: Dual quad-core processors; 8 GB RAM; 2 x 500 GB 7200 RPM SATA hard drives (as 500 GB RAID-1).

$200--$700 Uninterruptible Power Supply (UPS): If the computer room does not have a UPS (battery-backed, surge-suppressed power supply), or adding this server would exceed its load, budget to add a UPS for it. Budget $50-$100 to replace UPS batteries every few years.

$1500--$2500 Tape drive for backups: External SAS LTO-4 or LTO-5; budget about $25 per LTO-4 tape/$45 per LTO-5 tape for media. LTO-5 (but not LTO-4) supports LTFS (Linear Tape File System), which may aid tape usability. [LTO = Linear Tape Open, magnetic tape data storage.]

VT's preservation strategy is through membership in the MetaArchive Cooperative, which posts its membership fees and costs (e.g., server, storage space) at http://www.metaarchive.org/costs.

Who are the primary university personnel involved in or considered rudimentary to your ETD program and in what units do they work?

A computer science faculty member, the dean of the graduate school, and the director of the Digital Library and Archives shepherded the proposal through university governance that led to the ETD requirement.

Library staff who were involved in processing theses and dissertations came together to draft the online workflow for ETDs from submission to approval to access and storage to paying the ProQuest invoices. The DLA part-time programmer used this workflow to configure our automated computer processes. Now that we have a mature ETD program, the systems administrator is the key link in the continuum from submission to access to preservation. The hardware requires little time or attention, however, the aging locally designed scripts require some attention. Rewriting the original ETD-db software has taken over two years of part-time, largely student effort. However, the graduate school contracted to automate all of its manual procedures and forms, resulting in ETD submissions moving from the library to the graduate school in November 2012.

A Digital Technologies Librarians is spending six months part time on transitioning our legacy system to accommodate the graduate school's automated workflow. Shortly, library personnel will oversee the daily arrival of an ETD "pay load" and its deposit into the institutional repository, VTechWorks. Similarly, there are personnel costs associated with establishing the distributed preservation network participation (i.e., MetaArchive Cooperative) and writing plug-ins. But once instituted the time commitment to preservation is minimal but effective.

Because we have ETDs from 1995 forward, we also spend labor on occasionally redistilling PDFS (mostly those originating with LaTeX). We also handle requests for external access to VT-only ETDs (whether from our collection of scanned theses and dissertations or born-digital ETDs), which usually means locating and contacting authors for permission.

Graduate school personnel include 1 FTE for reviewing ETDs (4 staff) and their supervisor, the Director of Graduate Admissions and Academic Progress. The supervisor averages about one hour per week fielding questions, advising on "special" ETDs, handling ProQuest forms, etc. However, she also has a "speaker series" focused on ETDs, handles special ETD-related requests (e.g., early releasing), etc., which requires about five hours per year. She recently had special training provided by the VT Office of Export Control and more VT ETDs will be based on sponsored research. She and her staff spend 15-30 minutes reviewing each ETD and approximately 1,000 ETDs are approved each year, though the graduate school is continuing to grow its programs.

What other expenditures are associated with your institution's ETD program?

There are membership fees associated specifically and tangentially to ETDs. Virginia Tech is a founding member of the Networked Digital Library of Theses and Dissertations ($100/per year). Our preservation strategy requires membership in the MetaArchive Cooperative and the LOCKSS Alliance, but ETDs are a small percentage of our preserved collections.

How are resources (going to be) allocated to your institution's ETD program?

When VT switched to ETDs, we renamed the "binding fee" the "archiving fee." This fee has remained the same for over 30 years: $20. Dissertation students also pay the ProQuest fee of $25 so that their ETDs can be included in their various products. This fee goes through the library to ProQuest.

The library assumes all costs related to graduate school-approved ETDs. The graduate school assumes the costs for their staff and recently funded programming for automating many of their processes, including ETD submission and committee online approval.

In the past three years have your costs in the categories of technology, people, and other:

Increased? Decreased? Stayed roughly the same? And why?

Library costs increased because we were rewriting the ETD-db software at the same time that the graduate school contracted for their workflow software. After the library adapts the new workflow, which includes automating some currently manual processes, cost should again be minimal.

How are costs shaping the current planning and directions for your ETD program?

Little effect. We are migrating from our legacy system, which means receiving a daily "payload" from the graduate school's new automated workflow into our institutional repository, VTechWorks. The library and the graduate school are each absorbing their respective costs. The ProQuest fees and the archiving feel will remain the same.

Bibliography

Abrams, Stephen. 2012. "CDL: Cost Modeling." https://wiki.ucop.edu/display/Curation/Cost+Modeling.

Abrams, Stephen, Patricia Cruse, and John Kunze. 2012. "Total Cost of Preservation: Cost Modeling for Sustainable Services." https://wiki.ucop.edu/download/attachments/163610649/TCP-total-cost-of-preservation.pdf?version=4&modificationDate=1334613033000.

Ashley, K. 2000. "Digital Archive Costs: Facts and Fallacies." DLM Forum '99. http://ec.europa.eu/archives/ISPO/dlm/fulltext/full_ashl_en.htm.

AS 4390.1-1996 Australian Standard: Records Management.

"Australasian Digital Theses Program: Draft Business Plan 2006-2009." 2007. http://www.caul.edu.au/content/upload/files/adt/adt2006-2009businessplanV3.doc.

Ayris, Paul, Richard Davis, R. McLeod, R Miao, Helen Shenton, and Paul Wheatley. 2008. "LIFE2 Final Project Report". JISC. http://eprints.ucl.ac.uk/11758/.

Beagrie, Neil, Julia Chruszcz, and Brian Lavoie. 2008. "Keeping Research Data Safe: A Cost Model and Guidance for UK Universities." JISC. http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx.

Beagrie, Neil, Brian Lavoie, and Matthew Wollard. 2010. *Keeping Research Data Safe 2*. UK: JISC. http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Final report. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. 2010. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." http://brtf.scsd.edu.

Brown, Adrian. 2005. "Cost Modeling: The TNA Experience." The National Archives (UK). July 26, DCC/DPC joint Workshop on Cost Models. http://www.dpconline.org/docs/events/050726brown.pdf.

Chapman, Stephen. 2003. "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?" *Journal of Digital Information* 4 (2). http://journals.tdl.org/jodi/article/view/100.

"Cost Model for Digital Preservation." 2012. Accessed November 30. http://www.costmodelfordigitalpreservation.dk/.

Crespo, Arturo, and Hector Garcia Molina. 2001. "Cost-Driven Design for Archival Repositories". First ACM/IEEE-CS Joint Conference on Digital Libraries. http://www-db.stanford.edu/~crespo/publications/cost.pdf.

Davies, Richard. 2008. “Lifecycle Information for e-Literature (LIFE2) Final Project Report [Summary].” http://ie-repository.jisc.ac.uk/394/.

Davies, Richard; Paul Ayris, Rory McCleod, Hellen Sheton, Paul Wheatley. 2007. “How Much Does It Cost? The LIFE Project Costing Models for Digital Curation and Preservation.” *Liber Quarterly* 17 (3/4). http://liber.library.uu.nl/.

“Digital Preservation and Data Curation Costing and Cost Modelling.” 2012. http://wiki.opf-labs.org/display/CDP/Home.

Dugan, Robert E. 2002. “Information Technology Budgets and Costs: Do You Know What Your Information Technology Costs Each Year?” *Journal of Academic Librarianship* 28 (4) (July): 238–243. doi:10.1016/S0099-1333(02)00288-4.

Dutch National Archives. 2003. “From Digital Volatility to Digital Permanence: Preserving Databases.” http://www.ltu.se/cms_fs/1.83816!/file/Preserving_Databases.pdf.

Goldstein, Serge J., and Mark Ratliff. 2010. “DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data.” http://arks.princeton.edu/ark:/88435/dsp01w6634361k.

“Guide to Electronic Theses and Dissertations.” http://curric.dlib.vt.edu/wiki/index.php/ETD_Guide.

Harris, Carolyn L., Carol A. (Carol Ann) Mandel, and Robert A. Wolven. 1991. “Cost Model for Preservation: The Columbia University Libraries’ Approach.” *Library Resources & Technical Services* 35 (January): 33–54.

Hendley, Tom. 1998. “Comparison of Methods & Costs of Digital Preservation”. British Library Research and Innovation Centre.

Higgins, Sarah. 2008. “DCC Curation Lifecycle Model.” *International Journal of Digital Curation* 3 (1): 134–147. doi:10.2218/ijdc.v3i1.48. http://www.ijdc.net/index.php/ijdc/article/view/69.

Hole, Brian, Li Lin, Patrick McCann, and Paul Wheatley. 2010. “LIFE3: A Predictive Costing Tool for Digital Collections” presented at the iPres. http://www.life.ac.uk/3/docs/Ipres2010_life3_submitted.pdf.

“Institutions Requiring Electronic Thesis/dissertation Submissions.” https://docs.google.com/spreadsheet/ccc?key=0AtSglIhGWCkpdHJvOUNSZUZyRC04UXRUa0w3UmgtYWc&hl=en#gid=0.

Jewell, Christine. 1999. “Electronic Thesis and Dissertation System: Business Plan.” University of Waterloo Library Planning and Priorities Group. http://www.lib.uwaterloo.ca/ETD/business_plan.PDF.

Kalb, Sam. 2005. "QSpace – Queen's Institutional Repository Business Plan." http://dspace.ucalgary.ca/bitstream/1880/43365/3/q_space_planning_document.htm.

Kejser, Ulla Bøgvad; Anders Bo Nielsen; Alex Thirifays. 2009. "Cost Model for Digital Curation: Cost of Digital Migration" presented at the iPRES 2009: the Sixth International Conference on Preservation of Digital Objects. http://escholarship.org/uc/item/4d09c0bb.

Kenney, Ann. 2005. "Cornell Experience: arXiv.org." presented at the DCC/DPC Workshop on Cost Models for Preserving Digital Assets, July 26. http://www.dpconline.org/events/previous-events/137-cost-models.

Lazorchak, Butch. 2012. "Digital Asset Sustainability and Preservation Cost Bibliography | The Signal: Digital Preservation." http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/.

MacDonald, Ames R.W., and Daniel Yule. "Jarrow, Electronic Thesis, and Dissertation Software." http://journal.code4lib.org/articles/7486.

MacDonald, James R.W., and Yule, Daniel. 2012. "Jarrow, Electronic Thesis, and Dissertation Software." *Code4Lib Journal*. http://journal.code4lib.org/articles/7486.

National Archives of the Netherlands. 2005. "Digital Preservation Testbed Costs of Digital Preserving." Digital Preservation Testbed, Nationaal Archief [Netherlands]. http://en.nationaalarchief.nl/kennisbank/costs-of-digital-preservation-2005-0.

"NDLTD: Networked Digital Library of Theses and Dissertations." www.ndltd.org.

NDTLD. 2011. *Guide to Electronic Theses and Dissertations*. Blacksburg, Virginia: Digital Library and Research Laboratory, Virginia Tech. http://curric.dlib.vt.edu/wiki/index.php/ETD_Guide.

"Robert Gordon University Electronic Theses Business Case." http://www2.rgu.ac.uk/library/etds/business.html?CFID=49138276&CFTOKEN=85725408&jsessionid=5030bd7d591b78e222bb706f45165f4e3749TR.

Sanett, Shelby. 2002. "Toward Developing a Framework of Cost Elements for Preserving Authentic Electronic Records into Perpetuity." *College & Research Libraries* 63 (5) (September): 388–404. http://crl.acrl.org/content/63/5/388.abstract.

Shenton, Helen. "Life Cycle Collection Management." *Liber Quarterly* 13: 254–272.

Slats, Jacqueline, and Remco Verdegem. 2005. "Digital Preservation Cost Model." Nationaal Archief/Testbed Digitale Bewaring. http://en.nationaalarchief.nl/.

University of South Florida. Graduate School. 2013. "Frequently Asked ETD Questions [FSU]." http://www.grad.usf.edu/ETD-FAQ.php#f.

Wheatley, Paul. 2012. "Digital Preservation Cost Modelling: Where Did It All Go Wrong?" Accessed November 28. http://openplanetsfoundation.org/blogs/2012-06-29-digital-preservation-cost-modelling-where-did-it-all-go-wrong.

Zeller, Jean–Daniel. 2010. "Cost of Digital Archiving: Is There a Universal Model?" presented at the European Conference on Digital Archiving. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf.

8 Guide to Options for ETD Programs

Martin Halbert (University of North Texas)

Topics Covered

- Access Policies and Intellectual Property Issues.
- Deposit Procedures.
- Repository System Options.
- ETD Program Management.
- ETD Program Services.

8.1 Introduction

Many academic institutions around the world have implemented programs to store and manage electronic theses and dissertations (ETDs). The widespread growth of ETD programs and their perceived advantages for graduate education and long-term preservation and access to scholarship have been documented elsewhere (Lippincott 2010). This document is intended to provide useful guidance for academic decision-makers considering new implementations or overhaul of existing ETD programs, with a special focus on the options in such programs that may warrant special attention and discussion.[1] These options may need special attention for a variety of reasons. There may be many questions and factors to consider associated with particular choices in implementing ETD programs, either at the policy level or in particular aspects of implementation. Because ETD programs are still relatively new despite being widely implemented, there may simply be a lack of clear consensus or understanding among stakeholders about the fundamentals of such programs, what is meant or implied by the terminology used in planning ETD programs, or misunderstandings about assumptions (real or imagined) underlying ETD programs. While the benefits of ETD programs are now widely accepted, the range of options, pace of innovations in ETD services, and controversies surrounding different options in ETD programs may be justifiably daunting to academic administrators. Institutions may delay implementation or improvements to ETD programs because of fears and uncertainties over ETD program choices. Restricted or open access? Implement an ETD repository or lease a commercial service? Who will have responsibility for what functions? These are only some of the questions that must be considered and periodically re-considered when managing an ETD program.

[1] A great deal of useful research and information resources has now been produced concerning ETD programs that can provide well-informed advice and perspectives in planning for the many possibilities and choices to be made in depositing, accessing, and managing ETDs over time. Where there are citable resources, these will be referenced. Where there are no clear sources of information on a topic, an attempt will be made to descriptively set forth the different sides of the issue.

8.1.1 Benefits and Impacts of ETD Programs

The starting assumption for this briefing is that an institution perceives some value to implementing an ETD program, and that the decisions that require consideration are in the specifics of the implementation decision, rather than the fundamental question of whether or not such a program is worth implementing. This reflects the literature and practice within the field. While there have been some attempts at rigorous investigations into the value and impacts of ETD programs (Brown et al. 2010; Macduff 2009), most institutions that have implemented ETD programs do so because of a widespread perception that such programs have inherent (albeit hard to quantify) benefits in terms of improved management, preservation, and access to theses and dissertations. Seminal publications by Dr. Edward Fox and Gail McMillan have set forth the basic claims concerning the benefits of ETD programs. Annual conferences such as the international ETD Conference and the US ETD Association (USETDA) conference attest to the extensive interest and faith in ETD programs by many institutions. The benefits of ETD programs are often articulated directly in the institutional policies concerning such programs; a good example of this is the North Carolina State University ETD Guide (NCSU 2011). The basic value proposition of ETD programs is a relatively non-controversial question, and therefore lies beyond the scope of this briefing.

8.1.2 Key Decisions for ETD Programs

The more ambiguous aspects of ETD programs lie in the specifics of implementation options and decisions, and this is a more useful place to concentrate this discussion. The following are the key areas of decision-making that this document will address; these are the categories of high-level decisions that usually receive the most attention in planning an ETD program or overhauling it:

- Access policies.
- Deposit policies.
- Repository system options.
- ETD program management.
- ETD program services.

8.1.3 Information Resources on ETD Programs

When planning or considering the implementation of an ETD program, decision makers should know that there is now a significant body of research on this topic that may be consulted, and which has been collated in the project bibliography.[2] This briefing will call out debates from this literature that planners should be aware of, as well as referencing more detailed resources. Perhaps the most comprehensive clearinghouse of information on ETD programs is the international organization known as the Networked Digital Library of Theses and Dissertations (NDLTD). The NDLTD website provides a range of informative resources and planning documents at http://www.ndltd.org/resources to consult. This briefing was developed in consultation with the NDLTD leadership, and will reference many NDLTD resources.

[2] Available here, http://metaarchive.org/imls/index.php/IMLS_ETD_Project_Bibliography.

To understand the key choices in implementing ETD programs, recent descriptive surveys of such programs are a good place to start. A 2010 survey by Joan Lippincott of the members of the Coalition for Networked Information concerning ETD programs resulted in 88 responses from 142 institutions contacted. (Lippincott, 2010) This survey documented the widespread implementation of ETD programs and reported that the majority (73%) of responding institutions had already instituted an ETD program of some sort, with five additional institutions indicating that they were planning such an implementation. In 89% of the institutions, ETDs were reported to be a subset of larger institutional or consortial repository holdings. An implication of these figures is that institutions that are now considering ETD program implementations are likely to be in categories that the CNI survey did not focus on, including smaller institutions, or institutions that to date have had significant reservations about ETD programs. Wherever possible, this briefing will attempt to address the anticipated concerns of such institutions.

The CNI survey examined a range of factors and perceptions concerning ETD programs, including system implementation strategies, access and embargo considerations, format options, and other ETD services and policies to best serve graduate students. While these are certainly not the only decisions that must be made in planning an ETD program, these categories of options are a useful frame of context and will be used for the purposes of this briefing. The CNI survey is also not the only source that will inform this briefing; other reports will be cited as appropriate. Many of the following controversies are also informed from actual conversations reported anecdotally from the field by practitioners.

8.2 Access Policies and Intellectual Property Issues

Key issues

- Violation of copyright by authors of ETDs.
- Violation of copyright by those who access ETDs.
- Anxiety about quality of ETDs.
- Embargo controversies.
- Copyright and ownership of ETDs.

Perhaps the most intimidating set of options in ETD programs surrounds a complex series of inextricably linked issues associated with network access to and intellectual property rights of theses and dissertations. Curiously, these issues have only become widely visible and controversial as theses and dissertations have become digital and accessible via the Internet. While print theses and dissertations in most university library archives were traditionally made readily accessible to scholars seeking to access and read them, institutional decision-makers in recent years often have far more reluctant reactions to the idea of making ETDs freely accessible on the Web. When it manifests, this reluctance is perceived as inconsistent and somewhat ironic by many ETD program proponents. The response often runs as follows: Why should access to the electronic versions of theses and dissertations be restricted when access to print versions is not? As mentioned previously, one of the primary reasons for implementing ETD programs is the perceived value of making these key academic documents *more* accessible rather than *less* accessible. Given that increased awareness of a scholar's work is broadly perceived to be an advantage rather than a disadvantage in academia, improving the accessibility and discoverability of student theses and dissertations is seen as a benefit.

However, the anxieties that come forward concerning ETD accessibility usually revolve precisely around the tremendous increase in discoverability and reproducibility of documents that can be freely downloaded via the Web (the digital form of the electronic thesis or dissertation is usually conflated with Web accessibility when these concerns manifest, although these two properties of ETDs are obviously distinct). These anxieties usually sort out into the following areas of concern.

8.2.1 Violation of Copyright by Authors of ETDs

While the issue of copyright violation in theses and dissertations has always been an issue, it escalates when ETDs are made available on the Internet. Academic administrators may fear that copyright violations (intentional or unintentional) that may exist within student theses and dissertations will become more visible when exposed on the Web. The somewhat obvious rejoinder to this concern is that the institution should implement effective measures for detecting and preventing copyright violations whatever form the thesis or dissertation takes, whether electronic or print. There are now a growing number of tools for automatic detection of plagiarism, both commercial services and free options (although either option requires that someone in the institution does need to take responsibility for routine plagiarism checks). While it might be true that copyright violations are more prone to be discovered when the general discoverability of theses and dissertations increases, this does not seem to be a particularly compelling reason to limit access to ETDs.

A more rational, yet subtle, concern arises in situations in which limited permission has been obtained by the student from the copyright holder to reproduce an image or other content solely in the print copy of the thesis or dissertation submitted for graduation, but not in an openly accessible ETD. There are still categories of theses and dissertations for which this scenario is relatively common. Some disciplines in which this may be the case include art history and performance-based fields of study in which expectations have not caught up with the possibilities of the technology or institutional practice. One straightforward remedy for this scenario that has been put forward by institutions that seek to maximize accessibility of ETDs is simply education and awareness building for both students and their faculty committee members, with the aim that more general permissions be obtained for copyrighted material included in theses and dissertations such that ETDs can be made freely accessible. The other point is to again emphasize that effective measures for detecting and preventing copyright violations in ETDs should be in place procedurally.

8.2.2 Violation of Copyright by Those Who Access ETDs

Another controversy centers on fears that wider access to the theses and dissertations of an institution's students will make these works more likely to be plagiarized. While plagiarism is always something to be alert to, arguments to this effect again appear to be hollow upon closer consideration. While instances of plagiarism may (or may not) have increased because of widespread electronic access to theses and dissertations, that is not a reason to constrain access to the work of an institution's students. It may very well be another justification for implementing better mechanisms for automated mechanisms for detection of plagiarism, however. And it may also be another justification for awareness raising programs among students and faculty as to the specifics of what constitutes plagiarism and ways to prevent it.

8.2.3 Anxiety About Quality of ETDs

The most frequently raised argument against implementation of ETD programs is based on fear by either administrators or faculty that wider access to the theses and dissertations of an institution will somehow lead to greater exposure of deficiencies in these student works. This concern most often takes the form of reluctance expressed in closed meetings when the idea of an ETD program is initially being proposed, and is driven by (sometimes unarticulated) fears by administrators or faculty that their students' work is sub-par or somehow otherwise flawed and broader review by external reviewers will surface these faults.

Rejoinders to this viewpoint are clear upon closer inspection: if the administration or faculty have concerns about the quality of student work, or if there is a perception that faculty are allowing students to submit inadequate theses and dissertations, then steps should be taken to remediate this situation rather than seeking to obscure it. Of all the objections raised against providing electronic access to theses and dissertations this objection may be the most counter-productive as it may never actually be articulated, yet may be lurking silently behind–the-scenes so to speak. If an ETD program implementer suspects that this fear may be present, it is best to simply get it out in the open and discuss it objectively. Talking about such fears is usually the best way to overcome them.

The previous three controversies over ETD program access policies are relatively straightforward to respond to because they are based largely on misconceptions. The following access policy recommendations are more substantive.

8.2.4 Embargo Controversies

The most protracted discussions that occur in implementing ETD programs often involve complex discussions on various aspects of "embargos" in which access to ETDs is delayed temporarily or permanently. There is a tremendous range of embargo options, with the chapter on Access Levels and Embargoes more fully addressing the range of possibilities to consider. We will here only attempt to summarize the main categories of embargo arguments that are most often discussed during the creation of ETD programs.

Sometimes faculty members (especially in the humanities) are strong advocates of embargos. The claim has frequently been made that students who release a humanities dissertation through an ETD repository will be precluded from subsequently publishing their dissertation as a book. The argument is that such dissertations should therefore not be made publicly accessible and should be embargoed. The level of anxiety over this issue among some humanists is such that it sometimes leads to claims that all humanities ETDs should be permanently embargoed, or even that access to all institutional ETDs should be permanently embargoed by policy. Proposals such as these are generally considered extreme positions, and have been strongly contested in recent years for several reasons. A 2011 survey by McMillan et al. demonstrated that 72% of publishers will accept submissions of openly accessible ETDs for consideration, and an additional 14% indicated that they would consider such submissions if the contents and conclusions in the manuscript are substantially different from the ETD. In fact, only 4% of publishers responded that they would never consider such submissions. Some scholars report a significant advantage to releasing their ETD as an early version of a later publication, primarily because of the visibility and recognition it affords them in advance of the publication.

The opposite extreme position is that all ETDs must immediately be made permanently open access upon deposit. This position is countered by examples in which at least some period of embargo is strongly warranted. The strongest such objection to making a thesis or dissertation openly accessible immediately concerns the quite legitimate issues involved if there is an associated patent application. When there is a pending patent application based on some aspect of the work involved in a thesis or dissertation, many institutions constrain all public access to the work. This is because public release of any material that discloses substantive details of an invention before it is patented may likely constitute "prior art", and as such will render a patent application invalid. There are other situations in which an embargo of some period may be requested for good reasons; an example might be that the student is seeking to publish the dissertation with one of the tiny 4% minority of publishers that explicitly refuse to consider submissions of ETDs.

Most ETD programs have therefore resisted proposals for automatic blanket policies, but encourage open access while providing students with processes for requesting embargoes with appropriate justifications. These processes are often structured through the specifics of the deposit policy the institution implements (see the following section).

8.2.5 Copyright and Ownership of ETDs

The quite basic question of assertion of intellectual property for the ETD as a work is sometimes a source of disagreement or controversy. In most institutions the assumption is that the student author holds the copyright with the institutional ETD program retaining a permanent right to display the work in the ETD repository. This is not always the case; an example of an exception is MIT, which asserts copyright for theses and dissertations that are created with financial or technical support from the university (MIT, 2011).

Producing the thesis or dissertation may frequently be the first time the student has occasion to think of themselves as an author. This is another opportunity for an ETD program to usefully provide information concerning the rights of authors in asserting copyright over their work. The chapter *Guidelines for Implementing ETD Programs – Roles and Responsibilities* provides additional information about the kinds of instructional programs that an ETD program may wish to consider implementing in conjunction with other services.

8.3 Deposit Procedures

Key Issues

- Mandatory versus optional deposit.
- File formats.

After access and copyright issues, the next key area of decision in implementing an ETD program is the specification of a procedure for depositing the ETD. Questions concerning this aspect of ETD programs often get mired in very specific details, but should be informed by an overall set of goals for the ETD program.

8.3.1 Mandatory versus Optional Deposit

A fundamental question any ETD program must grapple with is whether or not all theses and dissertations are to be deposited into the institutional ETD repository. If there are exceptions, what rules structure the deposit process? When are ETDs made publicly accessible? Who is responsible for which roles in the entire lifecycle of the ETD?

In the 2010 CNI survey 43% of institutions reported that ETD deposit was mandatory for both doctoral and masters students. (Lippincott, 2010) Many institutions have transitioned to the viewpoint that the ETD submitted is the document of record, rather than the submitted print thesis or dissertation. Electronic theses and dissertations are potentially much easier to deposit, provide access to, and to preserve through replication.

8.3.2 File Formats

Questions often arise during implementation of ETD programs concerning which file formats should be allowed for ETDs. Many programs have concerns about the complexity of managing and preserving multiple file formats over long periods of time. This is again a complex topic, and this section will only attempt to highlight some of the relevant controversies; the chapter on *Managing the Lifecycle of ETDs: Curatorial Decisions and Practices* will more fully discuss this issue.

A 2008 survey by McMillan found that 85% of ETD programs accept PDF submissions of theses and dissertations, but there was far less consensus on other multimedia file formats. (McMillan 2008) While there is little consensus on which multimedia formats are the best to manage and preserve, there is a growing consensus that ETD programs must go beyond simple archiving of the thesis or dissertation as a PDF. Many theses and dissertations now include rich statistical datasets, images, or other associated information that comprise content associated with the intellectual work of the research project. An issue that has occupied many well-established ETD programs in the past few years has been the question of how to begin archiving these far more complex forms of information together with the basic PDF representing the thesis or dissertation submitted for degree candidacy. Both new and old ETD programs are taking up this issue as a capability that must be incorporated into ETD deposit workflows, but this is quite a difficult task. Checking the validity of incoming PDF submissions is difficult enough procedurally; how will thinly-staffed ETD programs take up the much more complicated task of validating and documenting complex ETD objects that may include dozens or even hundreds of distinct files? Creating effective metadata for such complex and variable ETD submissions is a daunting notion, yet virtually all experts agree that this level of metadata is a necessity for long-term management of ETDs. There is no such thing as benign neglect of objects in digital archives, which must be actively managed over time to survive at all. This issue warrants serious consideration by any ETD program, whether long-established or just coming on the drawing board, and there are no easy answers, only tradeoffs and priorities to consider.

8.4 Repository System Options

Key Issues

- Locally managed repositories.
- Commercially managed repositories.
- Consortial repositories and other hybrid Options.
- Dinstinction of access and preservation systems.

The 2010 CNI survey identified three main repository strategies for implementation of ETD programs: 1) locally managed repositories, 2) repositories managed by consortia, and 3) repositories managed by commercial firms. There are many perspectives on the pros and cons of these different options; the following is an attempt to summarize the salient points of these different positions.

8.4.1 Locally Managed Repositories

Also known as institutional repositories (IR), these are systems implemented at particular universities using open source software or locally developed systems. Many institutions implement a local system for managing a variety of institutional digital content (hence the name "institutional repositories") typified by ETDs, faculty scholarship, local grey literature, and locally digitized content. Such institutions realize operational efficiencies through managing all these different types of content in one functional system. There is now a large literature that describes the implementation of locally managed institutional repositories (Bailey 2011), detailing the pros and cons of individual systems and strategies for implementation. This brief guide will not attempt to summarize all of the options available to an academic library seeking to implement an institutional repository; but will instead categorize the broad types of options that should be considered when thinking about ETDs and IRs.

The first option is whether or not to limit the repository to ETDs. While the most common option is for ETDs to be managed as one type of content among many managed within the IR, there are some institutions that have long maintained a separate repository just for ETDs. A notable example of this standalone strategy is Virginia Tech who has been a leader in the ETD movement almost from the beginning. (Sharretts and French 1999) The advantage of this approach is focus and constraint; the repository does not need to accommodate any workflows and metadata beyond what is needed for acquisition and administration of ETDs. Because ETDs are the sole focus in this scenario, it may be possible and desirable to develop features that are specific to improving access and usability of ETDs. Examples include interfaces for browsing and studying ETDs within departments or network charts of students that studied under particular professors.

However, the far more common option is for the institutional repository to manage ETDs within a larger context of genres and content types. This strategy has many advantages. There are significant costs to maintaining a local IR, and centralizing the acquisition and management of all types of locally generated content can leverage staff and other resources required to sustain an IR operation. It may be far more affordable (or at least palatable) for an academic library's administration to fund one IR for everything (including ETDs), rather than separate systems and workflows for different type of content. Maintaining one centralized system for all locally generated content can also minimize the number of interfaces that users have to learn in order to access a library's locally generated content.

A question that receives a great deal of attention in implementations of local IRs is whether or not to locally install an open source software package (such as DSpace or Fedora), or instead outsource the IR to a commercial solution (such as Digital Commons). Commercial solutions will be discussed in the next section. Local implementations of open source software packages give the institution a great deal of control and independence, but require more up-front investment in the technical staff to run the system. Running a local IR system requires that specific technical staff be available locally to install and maintain the system. Managerial oversight must also be vested in someone so that ongoing decisions and oversight occur in a reasonable time frame. Smaller libraries often partner with their campus IT centers to run institutional repositories, but medium to large institutions usually invest in a library IT department that has the capability to run servers and manage software (most frequently in conjunction with integrated library catalog systems).

8.4.2 Commercially Managed Repositories

Outsourcing to a vendor can minimize the up-front implementation expenditures required for local staff to run a system, but will still require significant expenditures for the fees associated with leasing access to a commercial system. An outsourced option will also require that the workflows associated with the vendor's system be adopted locally. Despite the drawbacks, many libraries see outsourcing as an attractive option given the high cost of technical staff. Letting a large commercial vendor run the system is appealing in that running the software becomes someone else's problem, but it also requires a significant degree of confidence that the vendor will be able to resolve problems encountered in a satisfactory manner. The issue of eventual migration from one system to another also becomes somewhat more problematic in the case of a commercial vendor; once ETDs have been deposited for several years in a commercial system, it may be extremely difficult to extract them for use in another repository.

To the extent that the institution intends for the IR to function as its ETD repository, the decision about whether or not to implement a local IR or use a commercial service is a key decision to make, and the question comes down to priorities. What is the most important priority: flexibility and local control or minimization of technical staff?

8.4.3 Consortial Repositories and other Hybrid Options

It may be the case that there is a regional consortium that maintains a large shared repository system for ETDs and other varieties of locally generated digital content. An example is OhioLINK, a shared digital repository, including ETD collections, maintained for many institutions in Ohio. Another example includes the shared ETD repository services provided by the California Digital Library for UC system institutions. Leveraging investments in technical staff through a regional consortium can enable the creation of significantly more capable repositories and greatly enhance the IR capabilities of cooperating libraries.

8.4.4 Distinction of Access and Preservation Systems

An unobvious distinction that IR implementers should be cognizant of is the distinction between access and preservation capabilities in planning. Any ETD solution must distinguish these two functions. *Access* to deposited ETDs is the combination of user interfaces and programmatic training provided to users who wish to gain access to an institution's ETDs. *Preservation* of ETDs encompasses systems and negotiated organizational agreements that ensure that ETDs will be accessible over indefinite periods of time to future users. Preservation systems may be quite different from the IR installations that provide access to ETDs. When considering long-term preservation, implementers should consider strategies for distributing secure replicated copies of the ETDs in geographically dispersed locations, the classic strategy for long-term survival of print materials. This may entail inter-organizational partner agreements with other institutions such as the MetaArchive Cooperative facilitates. Thinking through options for ensuring long-term preservation of ETDs is a worthwhile investment of time and resources to protect the unique institutional assets that ETDs represent.

8.5 ETD Program Management

The first, and perhaps most fundamental, set of questions facing those who wish to implement an ETD program is *who will be responsible for what aspects of the program*? The roles and responsibilities among three key stakeholder categories must be thought through carefully: graduating student authors of ETDs, the graduate school, and the library. Students obviously produce the theses and dissertations, and must be clearly advised of what actions they must take to deposit the electronic versions of their work. Graduate schools are responsible for certifying that students have met all requirements for graduation, and must clearly articulate the specifics of these requirements. Libraries are responsible for long-term preservation of the ETDs, and (usually) maintenance of the ETD program.

Having stated these basics, there are obviously many specific procedures that must be developed if an ETD program is to succeed. Each of the three key stakeholders must understand what their responsibilities are, and what information the other stakeholders need to take action appropriately. Since libraries are the institutions responsible for long-term preservation of ETDs, libraries are most often the institutional actor that initiates an ETD program on a campus. But the students and graduate schools must also buy into the concept of an ETD program and become engaged early on if it is to succeed.

8.6 ETD Program Services

Some final options that should be considered in implementing ETD programs include the ancillary services that will be implemented as part of the program. If implemented well, these services can greatly enhance the utility of the program and help "sell" administrations on the justification for the program.

Advice concerning plagiarism is often an ancillary or associated service that goes hand-in-glove with ETD programs. By advising students and faculty on the specifics of what constitutes plagiarism and how to avoid it, an ETD program can significantly contribute toward the successful completion of a student's graduate program. If students and faculty advisors are better equipped to avoid plagiarism, more theses and dissertations will ultimately progress to completion.

Another type of ancillary ETD program service that can enhance graduate education overall are usage statistics and other indicators of ETD significance. Joan Lippincott highlights the ways that such reporting mechanisms can illuminate the impacts of particular ETDs, and the ways that ETD programs can broaden the visibility of the theses and dissertations of an institution. (Lippincott 2006)

8.7 Summary

There are many questions and factors to consider associated with particular choices in implementing ETD programs, either at the policy level or in particular aspects of implementation. The categories of high-level decisions that usually receive the most attention in planning an ETD program or overhauling it, include: Access Policies, Deposit Policies, Repository System Options, ETD Program Management, and ETD Program Services. In the area of access, this document addresses concerns around copyright, quality, embargoes and ownership ultimately underscoring that improving the accessibility and discoverability of student theses and dissertations is seen as a benefit. In the area of deposit, the document advocates for being informed by an overall set of goals for the ETD program, with respect to mandatory versus optional deposit, and file formats. In the area of repository systems, the document, rather than advocate for one approach, puts forward the pros and cons of each of locally managed, commercially managed, and consortial solutions. Finally the document emphasizes that each of the three key stakeholders, the authors, graduate schools and the library, must understand what their responsibilities are, and what information the other stakeholders need to take action appropriately.

Bibliography

Brown, Josh. 2010. “Literature Review of Research into Attitudes towards Electronic Theses and Dissertations (ETDs).” *UCL EPrints*. http://eprints.ucl.ac.uk/20424/.

Copeland, Susan. 2008. "Electronic Theses and Dissertations: Promoting 'Hidden' Research." *Policy Futures in Education* 6 (1): 87-96.

DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Fox, Edward A., Shahrooz Feizabadi, Joseph M. Moxley, and Christian R. Weisser, eds. 2004. *Electronic Theses and Dissertations: A Sourcebook for Educators, Students, and Librarians*. New York: Marcel Dekker.

Fyffe, Richard, and William C. Welburn. 2008. "ETDs, Scholarly Communication, and Campus Collaboration." *College & Research Libraries News* 69 (3): 152-155. http://www.ala.org/ala/mgrps/divs/acrl/publications/crlnews/2008/mar/etdsschcommcampucollab.cfm.

Lippincott, Joan K. 2006. "Institutional Strategies and Policies for Electronic Theses and Dissertations." *EDUCAUSE Center for Applied Research Bulletin (*13). http://net.educause.edu/ir/library/pdf/ERB0613.pdf.

Lippincott, Joan K., and Clifford A. Lynch. 2010. "ETDs and Graduate Education: Programs and Prospects." *Research Library Issues*, (270) (June): 6-15. http://publications.arl.org/rli270/7.

Massachusetts Institute of Technology. 2011. “Specifications for Thesis Preparation: 2010-2011.” MIT: Boston. http://libraries.mit.edu/archives/thesis-specs/thesis-specs-2010-2011.pdf.

McMillan, Gail. 2008. “ETD Preservation Survey Results.” *Proceedings of the 11th International Symposium on ETDs*, Robert Gordon University, Aberdeen, Scotland. (June) http://scholar.lib.vt.edu/staff/gailmac/ETDs2008PreservPaper.pdf.

McMillan, Gail, Marisa L. Ramirez, Joan Dalton, Max Read, and Nancy H. Seamans. 2011. “An Investigation of ETDs as Prior Publications: Findings from the 2011 NDLTD Publishers’ Survey.” Cape Town, South Africa. http://hdl.handle.net/10919/11338.

NDLTD. “ETD terms.” http://www.ndltd.org/resources/Definition_of_ETD_Terms_6_10_2010_NDLTD.pdf.

North Carolina State University. 2011. *Electronic Thesis & Dissertation (ETD) Guide, NCSU Graduate School Guidelines and Requirements for Master Theses and Doctoral Dissertations.* NCSU Graduate School; Raleigh, NC. http://www.ncsu.edu/grad/etd/docs/etd-guide.pdf.

Reeves, Sharon. "How to Set Up an ETD Submission Program." NDLTD Resources Website. http://www.ndltd.org/resources/how-to-set-up-an-etd-submission-program.

Glossary

The following glossary defines terms that may be unfamiliar to a reader. Where possible, we have based our definitions on the United States ETD Association (USetdA) glossary. Their complete glossary is found at http://www.usetda.org/wp-content/uploads/2013/04/ETD_Terms_and_Definitions_USETDA.pdf.

Aggregator – A service that harvests content or metadata from multiple organizations to provide another mode of access. ETD aggregators include national and international services like the NDTLD Union Catalog and state-based services like the Texas Digital Library.

Born-Digital – "An item is born-digital if it has been generated entirely electronically by using a word processor" and/or electronic hardware such as a digital camera. (USetdA)

Catalog – An organized collection of metadata about a content collection. Entries for ETD collections in a catalog can be at the collection-level or item-level.

Closed Access – "The full text and sometimes the metadata of closed access ETDs are only available to authorized members of University staff and external examiners for administrative purposes. This is also known as an "embargoed" or "No Access" ETD. This access condition is usually granted" for ETDs describing patent pending or proprietary technology, personally identifiable or sensitive data, or classified information. "Some universities allow a limited closed access restriction period to allow students time to publish journal articles or books from their ETDs." (USetdA)

Collection – "A specialized group of records in an institutional repository. ETD collections are common and are often the largest and initial collections in an IR." (USetdA)

Copyright – "A form of protection grounded in the US Constitution and granted by law for original works of authorship fixed in a tangible medium of expression. Copyright covers both published and unpublished works. As intellectual property law, copyright protects original works of authorship including literary, dramatic, musical, and artistic works, such as poetry, novels, movies, songs, computer software, and architecture. Copyright does not protect facts, ideas, systems, or methods of operation, although it may protect the way these things are expressed. Copyright is affixed to the author as soon as the work is fixed in any tangible form, and is not dependent on publication of the work. Authors may secure, and/or transfer all or a subset of their rights via a signed written agreement." (USetdA)

Crosswalk – A relationship of specific elements between different metadata standards. Crosswalks allow an organization to transform and store the same metadata in the schema most useful to a particular use case.

Dark Archive – A form of storage with no form of public access. Reasons for not providing access include IP and copyright restrictions. See Closed Access.

Deposit – "The electronic submission of an ETD. Usually an online process, the student logs in, is guided by a "wizard" of prompts and screens to provide metadata as well as upload of the ETD document file(s)." (USetdA)

Digital Curation – The management, preservation, and enrichment of digital resources.

Digital Preservation – "The management process of ensuring digital objects and information are accessible over the long term. Development of standards, format compatibility, format migration, and systems interoperability are important aspect of this process. Digital preservation systems are under development to provide appropriate digital preservation techniques." (USetdA)

Digitized ETDs – see retrospective digitization

Dim Archive – A form of online storage with access restricted to the ETD's original institution. This would include on campus users as well as those in the university's community who have access through off-campus signing.

Distribution License – A license agreement signed by ETD authors, "which grants certain rights to their institutions for making their works available to the public." Ideally, authors share their rights with their institutions. (USetdA)

Dublin Core – A metadata standard used to describe the basic features of an object such as author, title, document type, date, and rights. See Metadata.

Embargo – "Restricting access to an electronic document for a specific period of time. Also, called Publication Agreement. "Publication" is a technical term in legal contexts and especially important in copyright legislation. An author of a work generally is the initial owner of the copyright on the work. Copyrights granted to the author of a work include the exclusive right to publish and/or transfer rights to the work." (USetdA)

ETD – Electronic Thesis and Dissertation

Fair Use – The doctrine of fair use is codified in the US Code, Title 17, Sect. 107. "It sets out four factors to be weighed to determine whether a particular use is fair:

1. The purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes [transformation]
2. The nature of the copyrighted work
3. The amount and substantiality of the portion used in relation to the work as a whole
4. The effect of the use upon the current or potential market for the work

The distinction between fair use and infringement is not easily determined. There is no specific number of words that may safely be taken without permission. Acknowledging the source of the copyrighted material does not substitute for obtaining permission." (USetdA)

Fixity – The property that a digital object does not change over time. Fixity is checked with a variety of fixity information ranging from weak (filename and file size) to strong (cryptographic hash).

Graduate schools – Composed of deans of libraries, colleges, and graduate schools; assistant and/or associate deans; deans from various colleges or schools; and graduate school staff who handle many details surrounding ETD programs. In addition, this group includes two other important stakeholders: graduate students and graduate faculty, both of whom are intimately involved in the development of theses and dissertations.

Institutional administrators – Top-level decision makers such as the university president, provost, chief information officer, and representatives from graduate council and the office of general counsel. They are not involved in the day-to-day operation of an ETD program. Rather, they support the program in various ways, including providing general guidance and/or funding support. They may also be the links that ensure the cooperation among the various stakeholders.

Institutional Repository (IR) – "An online database that provides access to digital collections such as theses and dissertations for online viewing and provides the associated metadata regarding the documents (e.g. student and university name, year of graduation, document title, abstract, keywords). A type of digital repository designed to collect the work of a particular institution." (USetdA)

Intellectual Property Rights – "Intellectual property (IP) refers to creations of the mind: inventions; literary and artistic works; and symbols, names, images, and designs used in commerce. IP is divided into two categories: Industrial property, which includes inventions (patents), trademarks, industrial designs, and geographic indications of source; and Copyright, which includes literary and artistic works such as novels, poems and plays, films, musical works, artistic works such as drawings, paintings, photographs and sculptures, and architectural designs. Rights related to copyright include those of performing artists in their performances, producers of phonograms in their recordings, and those of broadcasters in their radio and television programs." (USetdA)

IT personnel – Information Technology workers are composed of roles such as chief information officers, systems administrators, program analysts, application specialists, computer support specialists, and help desk staff. IT personnel may be in a centralized university unit and/or the library's IT unit. They support the management of, among many other works, born-digital and retrospectively digitized theses and dissertations as ETD – related activities require running software applications and server hardware in a network environment.

Item Record – The catalog entry related to a single bibliographic entity, such as copy two of a thesis. In library lingo this is based on the barcode. A catalog entry can have several item records attached to it.

Light Archive – Online storage of digital objects that is accessible by users.

MARC (MAchine-Readable Cataloging) – A metadata standard used to record the description of another object and serves as the basis of many library catalogs. See Metadata.

Metadata – Information about an object. In this document, metadata refers to the elements that describe an object. Different types of information (descriptive, technical, etc) are recorded in different metadata standards. See Dublin Core, MARC, METS, PREMIS.

METS (Metadata Encoding and Transmission Standard) – A metadata standard used to record the relationships between digital objects such as the association of a dataset to the ETD it supports. See Metadata.

Microfilm – “An archival microform produced on 35mm film reels which contain micro-reproductions of documents for transmission, storage, reading, and printing. Microform images are commonly reduced about 25 times from the original document size.” (USetdA)

Open Access (OA) – “Information readily available on the Web at no cost to users and without access restrictions. Also, the scholarly communication reform movement that aims to make scholarly literature freely available on the Web.” “Open-access literature is digital, online, free of charge to users, and free of most copyright and licensing restrictions. The full text and metadata of open access ETDs are available for downloading and viewing by anybody with access to the World Wide Web.” (USetdA)

Patent – “A patent protects inventions or discoveries. Ideas and discoveries are not protected by the copyright law, although the way in which they are expressed may be. See Copyright.” (USetdA)

PREMIS (PREservation Metadata: Implementation Strategies) – A metadata standard used to record technical information about an object for preservation purposes such as file format, fixity information, and associated intellectual property rights. See Metadata.

Pre-print – “Documents in pre-publication status, such as a draft or version of an article, that have not yet been published, but may have been reviewed and accepted for publication; submitted but with no publication decision; or intended for publication and being circulated for comment.” (USetdA)

Publication – "In the broadest sense, publication is to make content available to the public. While specific use of the term may vary between country, it is usually applied to text, images, or other audio-visual content on any medium, including paper or electronic publishing forms such as websites, E-books, Compact Discs and MP3s. "Publication" is a technical term in legal contexts and especially important in copyright legislation.” (USetdA)

Redaction – Information that has been selectively withdrawn before publication. ETDs may have portions redacted for reasons such as intellectual property rights and academic fraud.

Retraction – A publication that has been entirely withdrawn. ETDs may be retracted for reasons such as intellectual property rights and academic fraud.

ROI (Return on Investment) – A metric to determine the efficacy of an investment to bring about a certain result. Potential ROIs for an ETD program include impact on institutional reputation and impact on author citation rate.

Restricted Access – “This generally signifies that the complete work or parts of the work will have access limited to a defined user community. During this time an ETD may be available only to the original home institution, although the metadata is generally available to the public. This term may be used to refer to ETDs that are available to a limited user community as well as ETDs where access is embargoed.” (USetdA)

Retrospective Digitization – “The digitization of print documents such as bound theses and dissertations. Digitization involves a scanning process, application of standards for images files, and may but does not necessarily include OCR (optical character recognition) conversion. Digitized collections may be image-based files and/or enhanced full-text files that have been subject to an OCR process.” (USetdA)

University-only Access – “The full text of university-only ETDs are only available to authorized members of the institution’s students, faculty and staff via login or IP [internet protocol address] restriction. Many universities allow interlibrary loan of university-access only ETDs. Sometimes referred to as ‘Campus-Only’ [i.e., IP] access.” (USetdA)

Version Control – The process of managing content as it changes due to edits, redactions, format migrations, and other processes.