Preserving Electronic Theses and Dissertations: Findings of the *Lifecycle Management for ETDs* Project

Martin Halbert
University of North Texas
1155 Union Circle #305190
Denton, TX, 76203
940-565-3025
martin.halbert@unt.edu

Katherine Skinner
Educopia Institute
1230 Peachtree Street
Atlanta, GA 30309
404-783-2534
katherine@metaarchive.org

Matt Schultz
MetaArchive Cooperative
1230 Peachtree Street
Atlanta, GA 30309
616-566-3204
matt.schultz@metaarchive.org

ABSTRACT
This paper conveys findings from four years of research conducted by the MetaArchive Cooperative, the Networked Digital Library of Theses and Dissertations (NDLTD), and the University of North Texas to investigate and document how academic institutions may best ensure that the electronic theses and dissertations they acquire from students today will be available to future researchers..

Categories and Subject Descriptors
E.1 [**Data Structures**]: *distributed data structures*. H.3.2 [Digital Libraries]: Information Storage, *file organization.* H.3.4 [Systems and Software]: *distributed systems*. H.3.6 [**Library Automation**]: *large text archives.* H.3.7 **[Digital Libraries**]: *collection, dissemination, standards, systems issues*.

General Terms
Management, Documentation, Performance, Design, Reliability, Standardization, Languages, Theory, Legal Aspects, Verification.

Keywords
Archival Information Packages, Data Management, Digital Archives, Digital Curation, Digital Libraries, Electronic Theses and Dissertations, ETDs, Digital Objects, Digital Preservation, Distributed Digital Preservation, Ingest, Interoperability, Micro-Services, Repository Software, Submission Information Packages.

1. INTRODUCTION
One of the most important emerging responsibilities for academic libraries is curatorial responsibility for electronic theses and dissertations (ETDs) which serve as the final research products created by new scholars to demonstrate their scholarly competence. These are important intellectual assets both to colleges and universities and their graduates. Because virtually all theses and dissertations are now created as digital products with new preservation and access characteristics, a movement toward ETD curation programs in both U.S. institutions and abroad began in the early 1990's and has continued to this day.

There are many articles documenting this movement. The Coalition for Networked Information (CNI) recently studied the history of ETDs and graduate education and conducted an international survey concerning ETDs that examined linkages between the growth of ETD programs, institutional repositories, open access and other important trends in higher education (Lippincott and Lynch, 2010). Additional key issues identified in the CNI survey are questions and uncertainty within institutions concerning ETD embargoes, ETD format considerations, costs of ETD programs, and the role of libraries in working with graduate schools to maximize benefits of ETD programs for students.

A basic point made by the CNI study and virtually all current literature on the ETD movement is that colleges and universities have been steadily transitioning from traditional paper/microfilm to digital submission, dissemination, and preservation processes. Increasingly, academic institutions worldwide are now accepting and archiving *only* electronic versions of their students' theses and dissertations, especially in archiving programs operated by academic libraries. While this steady transition in curatorial practice from print to digital theses and dissertations greatly enhances the current accessibility and sharing of graduate student research, it also raises grave long-term concerns about the potential ephemerality of these digital resources.

Our research focuses on answering the question: *How will institutions address the entire lifecycle of ETDs, ensuring that the electronic theses and dissertations they acquire from students today will be available to future researchers?* 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), as well as in the more detailed senses of the digital lifecycle management model as articulated by the Digital Curation Centre in the UK (Higgins, 2008). A key outcome of our research and documentation will be a clearly articulated lifecycle model specific for ETDs.

In order to unpack this complex issue and to assess the library field's ETD lifecycle-management needs and practices, leaders of the Networked Digital Library of Theses and Dissertations (NDLTD) and the MetaArchive Cooperative conducted a series of investigations during 2008-2010. These efforts included surveys, a pilot project, and meetings of the leadership of the two groups, each of which are concerned with different aspects of preserving ETDs. The research team then embarked upon a US Institute for Museum and Library Services-funded project in 2011 to develop guidelines for ETD lifecycle management, software tools to facilitate ETD curation, and educational materials to help prepare ETD curators. As one component of this project, we conducted a focus group with stakeholders. We describe our findings from these surveys below.

1.1 Surveys of ETD Curation Practices

In order to assess practitioner needs and the current status of the field, the MetaArchive Cooperative and the NDLTD conducted a survey in 2007/2008 to examine ETD practices and associated concerns in institutions either currently engaged in ETD programs or considering such preservation service programs. The on-line survey was distributed through five major listservs and received 96 responses, primarily from academic institutions that were providing or strongly considering collection of ETDs and associated ETD services (McMillan, 2008).

Of the survey respondents, 80% accept ETDs, and 40% accept *only* ETDs. The ETD programs report that they accept many formats (more than 20) beyond PDF documents, including images (92%), applications (89%), audio (79%), text (64%) and video (52%). The average size of these programs was 41 GB, and respondents reported 4.5 GB/year average growth. We found that the repository structures used by respondents also vary widely. The more popular approaches included locally developed solutions (34%), DSpace (31%), ETD-db (15%), and such vendor-based repositories as bepress (6%), DigiTool (6%), ProQuest (6%), and CONTENTdm (6%).

This diversity of system types—presumably at least somewhat representative of the overall industry—presents an array of challenges for preservation. Each of these repository systems requires preservation attention during the ingest process to ensure that the materials are submitted in such a way that it is possible to retrieve them and repopulate that repository system with the content. This demands that content carries with it a level of context, and that context differs across repository structures.

The digital collections file and folder structures used by respondents also varied widely. Most respondents reported that their ETD collections are not structured in logically named, manageable virtual clusters. In fact, more than a quarter of respondents reported that their ETD collections are stored in one mass upload directory. This raises many preservation readiness challenges. How can the institution preserve a moving, constantly growing target? How can they ensure that embargoed and non-embargoed materials that often co-exist in the same folder are dealt with appropriately? How will the institution know what these files are if they need to repopulate their repository with them, particularly if they are stored in a repository system that does not elegantly package metadata context with content at export? Only 26% of the institutions manage their ETD collections in annual units. Another 26% use names (departments, authors) or disciplines as unit labels. Seven percent reported using access level labels and another 13% did not know.

The survey also collected information about what information institutions would need to make decisions concerning ETD preservation programs. Perhaps the most remarkable finding from this survey was *that 72% of responding institutions reported that they had no preservation plan for the ETDs they were collecting*.

The responses to this survey led the same researchers to conduct a follow-on survey in 2009 that probed more deeply into digital preservation practices and concerns (Skinner and McMillan, 2009). This survey included questions concerning institutional policies, knowledge and skills needed for digital preservation activities, level of desire for external guidance and expertise in digital preservation, and perceptions about relative threat levels of different factors in the long-term survivability of digital content.

Based on these findings, the MetaArchive Cooperative and the NDLTD undertook a joint pilot project in 2008-2010 to further explore and understand issues highlighted in the surveys and to respond to concerns of their respective memberships about preservation of ETDs. In the course of this pilot project, a group of institutions that are members of both organizations (including Virginia Tech, Rice University, Boston College, and others) worked together to discuss, analyze, and undertake experiments in different aspects of lifecycle management of ETDs, and to identify problem areas experienced by multiple institutions. The pilot project group also explored the literature to better understand what has been published to date on different digital lifecycle management topics, and how such publications relate to ETDs.

During this pilot project, as another means of assessing needs, Gail McMillan (NDLTD) and Martin Halbert (MetaArchive Cooperative) asked a large number of ETD program leaders about their concerns about ETD lifecycle management during workshops conducted at each of three annual ETD conferences hosted by the NDLTD from 2008-2010. Findings from the pilot project analysis and workshop inquiries were reviewed and discussed at three joint planning meetings of the NDLTD board and MetaArchive leadership during this period. They were consistent with the initial findings of the 2007-8 ETD survey.

Similarly, as the *Lifecycle Management for ETDs* project kicked off in 2012, the research team hosted a focus group in conjunction with the February Texas Electronic Theses and Dissertations Association meeting in Denton, Texas. Respondents in this focus group included both College of Arts and Sciences representatives and library representatives. The concerns raised by this group mirrored our earlier findings—most are involved in ETD programs and are either already electronic *only* or will be in the near future. The collection structures, file-types accepted, and repository infrastructures vary wildly. All attendees agreed that establishing documentation, tools, and educational materials that encourage better, more consistent ETD curatorial practices are of great need and should be of value to virtually all categories of academic institutions within the United States and internationally.

2. GUIDANCE DOCUMENTS

There is need for guidance documents in a variety of specific ETD lifecycle management topics to advance the capabilities of institutions that administer ETD service programs. The *Lifecycle Management for ETDs* project has worked to fill these gaps. The research team strongly feels that as a field we need to better understand, document, and address the challenges presented in managing the entire lifecycle of ETDs in order to ensure that colleges and universities have the requisite knowledge to properly curate these new collections. The research team has developed draft documentation on a number of topical areas, as briefly described below.

2.1 Introduction to ETDs

Prepared by Dr. Katherine Skinner and Matt Schultz (Educopia, MetaArchive), this document introduces the "Guidelines" and chronicles the history of ETDs. Using survey data and research findings, it describes the evolving and maturing set of practices in this area. It discusses the philosophical and political issues that arise in this genre of content, including what to do with digitized vs. born-digital objects, how to make decisions about outsourcing, and how to deal with concerns about future publications and

embargoed materials in the lifecycle management framework. The chapter provides a conceptual overview of a lifecycle model for ETDs that makes direct connections between the model and the individual guidance documents described below.

2.2 Access Levels and Embargoes

Prepared by Geneva Henry (Rice University), this document provides information about the ramifications of campus policy decisions for or against different kinds of access restrictions. It defines access restriction and embargo, and discusses reasons for each, including publishing concerns, sensitivity of data, research sponsor restrictions, and patent concerns. It discusses how institutions may provide consistent policies in this area and how policies might impact an institution's lifecycle management practices. It also reviews and compares existing university policies and makes policy recommendations.

2.3 Copyright Issues and Fair Use

Patricia Hswe (Penn State) chronicles ETD copyright and fair use issues that arise both in the retrospective digitization and the born-digital acquisition of theses and dissertations. It discusses institutional stances and guidelines for sponsored research and student work, and also reviews copyright and fair use issues with respect to commercial publishers (including e-book publishers) and vendors such as ProQuest. It seeks to provide clarifying information concerning publisher concerns and issues, providing a concise summary of the relevant information for stakeholders.

2.4 Implementation: Roles & Responsibilities

Xiaocan (Lucy) Wang (Indiana State University) documents the variety of stakeholders who impact and are impacted by the transition to electronic submission, access, and preservation of theses and dissertations, including such internal stakeholders as institutional administration (e.g., president, provost, CIO, general counsel), graduate schools (administrators, students, faculty), libraries (administrators, digital initiatives/systems divisions, technical services, reference), and offices of information technology, and such external stakeholders as commercial vendors/publishers, NDLTD, access harvesters (e.g., OCLC), and digital preservation service providers (e.g., MetaArchive, FCLA, DuraCloud). It emphasizes the range of functions played by these stakeholders in different management phases and institutions.

2.5 Demonstrations of Value

Dr. Yan Han (University of Arizona) provides guidance for institutions concerning assessment of ETD usage, and how communicating such assessment metrics can demonstrate a program's benefits to stakeholders. Han also documents practical examples of documenting and conveying usage metrics for stakeholder audiences, including the university, the students, and the research community more generally. He provides practical guidance for collecting, evaluating, and interpreting usage metrics in support of ETD programs, and discusses how it may be used to refine and promote this collections area.

2.6 Formats and Migration Scenarios

What factors should be considered by colleges and universities to determine what formats they should accept? How can they manage on an ongoing basis the increasingly complex ETDs that are now being produced by students? Bill Donovan (Boston College) discusses these format issues, including "data wrangling" practices for legacy content and migration scenarios for simple and complex digital objects in ETD collections.

2.7 PREMIS Metadata and Lifecycle Events

Another issue revealed in the needs assessment process was that most institutions do not have workflows and systems in place to capture the appropriate levels of metadata needed to manage ETDs over their entire lifecyle,. Daniel Alemneh (University of North Texas) informs stakeholders and decision makers about the critical issues to be aware of in gathering and maintaining preservation metadata for ETDs, not just at the point of ingestion, but subsequently, as ETDs often have transitional events in their lifecyle (embargo releases, redactions, etc.). This guidance document will both inform and reinforce the software tools around PREMIS metadata that we are building.

2.8 Cost Estimation and Planning

Gail McMillan (Virginia Tech) provides institutions with information on costs and planning, laying out the critical paths that many ETD programs have charted to date. This document provides cost-benefit analyses of multiple scenarios to give institutions a range of options to consider for their local needs.

2.9 Options for ETD Programs

Our surveys and focus group have demonstrated that many institutions are delayed in ETD program planning simply because they do not have a clear understanding of the range of options to consider in implementing an ETD program. Restricted or open access? Implement an ETD repository or lease a commercial service? Who has responsibility for what functions? Dr. Martin Halbert (University of North Texas) explains the relevant decisions institutions must make as they set up an ETD program and clarifies the pros and cons of different options.

3. LIFECYCLE MANAGEMENT TOOLS

The research team is developing and openly disseminating a set of software tools to address specific needs in managing ETDs throughout their lifecycle. These tools are modular micro-services, i.e. single function standalone services that that can be used alone or incorporated into larger repository systems. Micro-services for digital curation functions are a relatively new approach to system integration pioneered by the California Digital Library and the Library of Congress, and subsequently adopted by the University of North Texas, Chronopolis, MetaArchive, Archivematica, and other digital preservation repositories.

The micro-services described below draw upon other existing open source software tools to accomplish their aims. The intent of creating these four micro-services is that they will catalytically enhance existing repository systems being used for ETDs, which often lack simple mechanisms for these functions.

3.1 ETD Format Recognition Service

Accurate identification of ETD component format types is an important step in the ingestion process, especially as ETDs become more complex. This micro-service will: 1) Enable batch identification of ETD files through integration of function calls from the JHOVE2 and DROID format identification toolkits; and 2) Structure micro-service output in ad hoc tabular formats for importation into repository systems used for ETDs such as DSpace, and the ETD-db software, as well preservation repository software such as iRODS and DAITSS and preservation network software such as LOCKSS.

Components & Basic Requirements:

JHOVE2, DROID, XML output schema, Utility scripts (run commands, output parsers, etc.) & code libraries, API function calls, System requirements, Documentation & instructions

3.2 PREMIS Metadata Event Record-keeping

One gap highlighted in the needs analysis was the lack of simple PREMIS metadata and event record keeping tools for ETDs. This micro-service needs to: 1) Generate PREMIS Event semantic units to track a set of transitions in the lifecycle of particular ETDs using parameter calls to the micro-service; and 2) Provide profile conformance options and documentation on how to use the metadata in different ETD repository systems.

Components & Basic Requirements:

PREMIS Event profiles (example records) for ETDs, Event-type identifier schemes and authority control, AtomPub service document & feed elements, Utility scripts (modules) & code libraries, API function calls, Simple database schema & config, System requirements, Documentation

3.3 Virus Checking

Virus checking is an obvious service needed in ETD programs, as students' work is often infected unintentionally with computer viruses. This micro-service will: 1) Provide the capability to check ETD component files using the ClamAV open source email gateway virus checking software; 2) Record results of scans using the PREMIS metadata event tracking service; and3) Be designed such that other anti-virus tools can be called with it.

Components & Basic Requirements:

ClamAV, Utility scripts (run commands, output parser, etc.) & code libraries, API function calls, System requirements, Documentation & instructions

3.4 Digital Drop Box with Metadata Submission Functionality

This micro-service addresses a frequently sought function to provide a simple capability for users to deposit ETDs into a remote location via a webform that gathers requisite submission information requested by the ETD program. The submission information will: 1) Generate PREMIS metadata for the ETD files deposited; 2) Have the capacity to replicate the deposited content securely upon ingest into additional locations by calling other Unix tools such as rsync; and 3) Record this replication in the PREMIS metadata.

Components & Basic Requirements:

Metadata submission profile(s), Client/server architecture, GUI interface, SSL, authentication support, Versioning support, Various executables, scripts & code libraries, Database schema & config, System requirements, Documentation

All of these tools will be documented and released in 2013 via the project site: http://metaarchive.org/imls.

4. CONCLUSIONS

The first phase of this project has helped to reinforce preliminary research we had conducted regarding ETD lifecycle management practices (or the significant lack thereof). The field has a dire need for descriptive, not proscriptive, documentation regarding the range of ETD programs that institutions have designed and implemented to date, and the variety of philosophical, organizational, technical, and legal issues that are embedded therein. The field also has a stated need for lightweight tools that can be quickly implemented in a range of production environments to assist with some of the commonly needed curatorial practices for lifecycle management of these collections.

5. ACKNOWLEDGMENTS

We greatly appreciate the generous support of the Institute for Museum and Library Services (IMLS).

6. REFERENCES

Caplan, Priscilla. "The Preservation of Digital Materials." *Library Technology Reports*, (2008) 44, no. 2.

Conway, Paul. "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly*, (2010) 80:1, 61-79.

Fox, Edward A., Shahrooz Feizabadi, Joseph M. Moxley, and Christian R. Weisser, eds. *Electronic Theses and Dissertations: A Sourcebook for Educators, Students, and Librarians*. New York: Marcel Dekker, 2004.

Halbert, Martin, Katherine Skinner and Gail McMillan. "Avoiding the Calf-Path: Digital Preservation Readiness for Growing Collections and Distributed Preservation Networks," *Archiving 2009*, Arlington, VA, May 2009, p. 86-91.

Halbert, Martin, Katherine Skinner and Gail McMillan. "Getting ETDs off the Calf-Path" ETD 2009: *Bridging the Knowledge Divide*, Pittsburgh, PA, June 10-13, 2009. Sharon Reeves, ed. http://conferences.library.pitt.edu/ocs/viewabstract.php?id=733&cf=7

Hall, Susan L., Lona Hoover, and Robert E. Wolverton, Jr.. "Administration of Electronic Theses/Dissertations Programs: A Survey of U.S. Institutions." Technical Services Quarterly 22, no. 3 (2005): 1-17.

Lippincott, Joan K. "Institutional Strategies and Policies for Electronic Theses and Dissertations." *EDUCAUSE Center for Applied Research Bulletin*, no. 13 (2006). http://net.educause.edu/ir/library/pdf/ERB0613.pdf

Lippincott, Joan K., and Clifford A. Lynch. "ETDs and Graduate Education: Programs and Prospects." *Research Library Issues*, no. 270 (June 2010): 6-15. http://publications.arl.org/rli270/7

McMillan, Gail. "ETD Preservation Survey Results." *Proceedings of the 11th International Symposium on ETDs*, Robert Gordon University, Aberdeen, Scotland. (June 2008)

http://scholar.lib.vt.edu/staff/gailmac/ETDs2008PreservPaper.pdf

McMillan, Gail, and Katherine Skinner. (2010) "NDLTD/MetaArchive Preservation Strategy." (3rd ed.) http://scholar.lib.vt.edu/theses/preservation/NDLTDPreservationPlan2010.pdf

Skinner, Katherine, and Gail McMillan. "Surveys of Digital Preservation Practices and Priorities in Cultural Memory Organizations." 2009 NDIIPP Partners Meeting, Washington, D.C., June 24, 2009.

http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp09/docs/June24/NDIIPP_Partners_2009_finalRev2.ppt