# A Practical, Working and Replicable Approach to ETD Preservation

**Catherine M. Jannik**
Digital Initiatives Manager, Library and Information Center, Georgia Institute of Technology, Atlanta, GA

**Robert H. McDonald**
Associate Director of Libraries for Technology and Research, Florida State University, Tallahassee, FL

**Gail McMillan**
Director, Digital Library and Archives, University Libraries, Virginia Polytechnic Institute and State University, Blacksburg, VA

*Keywords:* ETD Preservation, Preservation Networks, Digital Preservation, Preservation Collaborations.

## ABSTRACT

Three NDLTD member institutions have created a cooperative digital preservation network for ETDs. This dark archive replicates and stores for emergency restoration their own and each of the other institutions' ETDs using open source software and off-the-shelf equipment. We propose it as a model to be replicated on each of the NDLTD member continents for global preservation purposes and as a means to create a community-centered on the digital preservation of ETDs.

Our university libraries are involved in two related digital preservation projects. Along with three additional universities and the U.S. Library of Congress, we are creating a distributed digital preservation network for critical and at-risk content called the MetaArchive of Southern Digital Culture (http://www.metaarchive.org). This network is part of the LC National Digital Information Infrastructure and Preservation Program (http://www.digitalpreservation.gov/) that is charged with providing, "a national focus on policy, standards and technical components necessary to preserve digital content." Simultaneously, the Association of Southeastern Research Libraries is pilot testing a similar ETD preservation strategy.

Our presentation will describe the accomplishments to-date, including the development of the distributed preservation network infrastructure based on the modified LOCKSS (Lots of Copies Keep Stuff Safe) software. Practical issues that we will also cover include: cost and specifications, metadata issues for digital preservation, as well as specific modifications for caching the institutional ETD repositories involved in these projects.

## 1. INTRODUCTION

There is a critical need at the national and international levels for more effective models of inter-institutional cooperation in digitally preserving collections and providing access to such materials, including electronic theses and dissertations. While many projects have developed technologies and standards for isolated preservation efforts at individual institutions, there have been far fewer projects that mobilized ongoing efforts involving large numbers of institutions spanning the globe.

Our universities are involved in two digital preservation projects. The MetaArchive of Southern Digital Culture addresses the preservation of the rapidly expanding body of humanities scholarship produced in digital formats. This project aims to understand and advance our shared capabilities to preserve such collections and provide access to their contents. We are also involved in a second project specifically about sharing preservation responsibilities for ETDs among members of the Association of Southeastern Research Libraries (ASERL).

The scope of this preservation need is captured in *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation*, which was jointly sponsored by the National Science Foundation and the Library of Congress. This report states, "…from a long-term preservation perspective, there is a dark side to the rapid growth in digital information. The technologies, strategies, methodologies, and resources needed to manage digital information for the long term have not kept pace with innovations in the creation and capture of digital information." (Hedstrom 2003. p. vii) "Libraries, archives, museums, and other cultural institutions that have preservation as part of their core mission need solutions to digital preservation challenges if they are to play a meaningful role in preserving our intellectual and cultural heritage." (Hedstrom 2003. p. viii)

What is fundamentally at stake is nothing less than our ability as a culture to preserve digital collections in the 21st century. Without the incremental advances of projects such as the MetaArchive, we risk entering a different kind of Dark Age in which substantial portions of the cultural record will be lost to posterity.

### 1.1 Distributed Archiving Strategies

It is generally accepted that most effective digital preservation practices will succeed through some strategy for dispersing copies of content in secure, distributed locations over time. Centralized approaches to digital preservation can be dismissed as inadequate due to the inherent shortcomings of current technology and long understood disadvantages of reliance on single instances of profoundly important materials custodians of rare books and manuscripts have dealt with for centuries.

Current digital preservation practice most often consists primarily of geographically and institutionally homogeneous replication of content by the host institution. This approach leaves that content at the mercy of that particular institution's possible technical infrastructure anomalies and vulnerable to destruction through both man-made and natural disaster. A network of geographically and institutionally diverse digital repositories adhering to best practices, such as those set forth in the Research Libraries Group's *Attributes of Trusted Digital Repositories* (RLG 2002) and the *Reference Model for an Open Archival Information System* (CCSDS 2002) ISO standard, eliminates many of these threats.

By basing our approach to digital preservation on leading preservation software for distributed digital replication (Lots of Copies Keep Stuff Safe, LOCKSS), the MetaArchive and ASERL

preservation networks establish from the beginning a distributed means of replicated archives. This approach addresses the geographic and institutional heterogeneity required to safeguard each institution's digital resources. In addition to creating the network and adapting the LOCKSS software for use in this manner, the MetaArchive preservation network is concerned with providing model partnership agreements and establishing best practices other institutions and organizations can use as a blueprint for emulating the preservation network. An overarching goal of the MetaArchive consortium members is to encourage others, including the NDLTD, to use this type of distributed preservation. In order to make the network easily replicable, MetaArchive and ASERL utilize open-source software and off-the-shelf hardware.

The progress made thus far by the partners lays a firm foundation for advancing digital preservation. The MetaArchive already in place is poised and eager to further the maturation of techniques needed for automated format migrations in collaborative digital archives, develop improved mechanisms for digital content ingestion from a variety of repositories, and design and prototype automated means for sharing and dissemination of metadata records for digital items. The MetaArchive includes ETD collections. However, the seven-university ASERL project emulates the MetaArchive network solely for ETDs. Since the MetaArchive project is more mature, this presentation focuses on it. However, progress on the ASERL ETD preservation project will also be reported at the Sydney conference.

## 2. METAARCHIVE APPROACH TO PRESERVATION NETWORKS

The approach to preservation networks that is being implemented currently by the MetaArchive partners is that of distributed redundant data storage at a bit stream level. This current mechanism does not include an end-user access mechanism but rather one for institutional access to preserved items. The redundancy is spread out over six different institutions utilizing the backbone of the Internet2 Abilene Network and the local connections of the Florida Lambda Rail, the Southern Crossroads (SoX) network consortium, and the Mid-Atlantic Crossroads (MAX) network consortium (Internet2 2005). The geographic area extends between Florida, Alabama, Georgia, Kentucky and Virginia and spans more than 200,000 square miles. The process for ingesting the digital material and storing it across the different server nodes is automated and managed via the LOCKSS Software Architecture. MetaArchive is a private dark archive of digital objects and is only open to the partners' servers. Thus if any one server node fails, it can be restored either from the primary source of the digital object or from any of the other five server nodes. The restoration process would be automatic once the server node was restored to an online status and could reconnect with the network.

### 2.1. Technical Infrastructure

It is impossible to speak of the MetaArchive technical infrastructure without a cursory glance at the overall hierarchy of the preservation collaborative. This is best done by looking at the components of the network via the OAIS reference model (CCSDS 2002). The MetaArchive framework is comprised of four layers: 1) Consortial Administration 2. Archival Storage 3) Content Ingestion and Replication (LOCKSS Software and Hardware and Network Connectivity) and 4) Shared Collection Description. Of these layers 1, 2 and 3 align directly with functions of the OAIS model while layer 4 is analogous to components of the OAIS Data Management layer.

**Consortial Administration** at this point is completely driven by human administration and oversight and is analogous to the Administrative layer of the OAIS model. This layer provides the common meeting point for the decisions made about content ingestion, archival planning, and inter-institutional long-term storage and access agreements. This work has been done with

group meetings in Atlanta at Emory University, video conferencing across the I2 Commons hosted by Florida State University, phone conferencing using AT&T conferencing services hosted by Emory University, and Web collaboration software (e.g., MoinMoin Wiki, PHP Collab (Project Management Software), and iVocalize Chat/VOIP meeting room).

**Archival Storage Layer** is the OAIS layer that will be investigated least by this project. MetaArchive will at a later date create a bridge between the content ingestion and replication system with a modular component that will enforce archival storage (format migration or emulation and another layer of data integrity checks that is informed by or communicates with the integrity checks contained within the LOCKSS software).

**Content Ingestion and Replication** is the layer that will be described further. It consists of the LOCKSS software ingestion and replication component as well as a hardware component that makes use of Linux systems administration tools for allocating disc space among partner nodes.

**Shared Collection Description** is the infrastructure that was developed to enable and manage collection description for the purposes of identifying *at risk* digital content among institutional partners. The Shared Collection Description layer feeds directly into the manifest page (i.e., permission statement) and archival unit hierarchy contained within the LOCKSS software component and is key to any future Data Management layer.

### 2.2. LOCKSS

Core to the MetaArchive Technical Architecture is the use of the LOCKSS software platform. The installation that is currently running in the test network is completely closed except to the nodes housed at the partner institutions. This diverges from the standard LOCKSS installation and allows more control over where digital content is stored and accessed and it enables the local MetaArchive partners to set policies on access and control based on local institutional priorities and does not require LOCKSS caches outside of the MetaArchive Network to function. This also limits the number of copies of a work that are available. However, the partnership does intend to expand its six node network at the end of our research and development stage of the project.

In its current stage, MetaArchive is only developing strategies for bit preservation and access (Reich & Rosenthal 2001). Future endeavours would include strategies for an archival storage layer or a data use migration layer (Rosenthal et al. 2005). Current bit preservation and access is accomplished using automated MD5 checksums along with the LOCKSS polling algorithm that checks each host node's content against each other for faults, additions, or subtractions. This decentralized model does not necessarily preserve bits within the framework of an individual institution's access but provides a cost sharing model for bit preservation and access within the MetaArchive partnership.

## 3. ETD PRESERVATION

Various ETD models have evolved, including commercial (e.g., ProQuest), consortia (e.g., OhioLink), institutional (e.g., Electronic Theses at Robert Gordon University Library), national (e.g., Australian Digital Theses), and international (e.g., NDLTD). A review of the literature brought to the fore several previously unknown ETD initiatives. However, they lack specific digital preservation components while the MetaArchive and ASERL projects are the only concerted ETD preservation initiatives. Below are examples of national and international ETD initiatives directly relevant to the NDLTD.

There are 1463 Arabic and French language ETDs in Algeria (www.dst.cerist.dz/equipe-edition-electronique.htm). The original floppy disks and CD-ROMs are stored in boxes and cupboards (Bakelli and Benrahmoun 2003 p. 2) mostly in the main offices of CERIST in Algiers, though after the library converts ETD documents to PDF, the files for each language are stored on a separate server. About one-third of the collection is being used in three "preservation" experiments: (1) determining the best among various file formats; (2) refreshing media chosen over migration or emulation; and (3) reformatting ETDs into XML. None, however, address long-term access to this body of Algerian works.

DATAD: Database for African Theses and Dissertations is sponsored by the Association of African Universities, with an emphasis on raising the worldwide profile and accessibility of research by African scholars. While the American Center for Research Libraries is involved, issues of preservation have not yet been addressed, the focus being on intellectual property, governance, and dissemination. (DATAD 2004) Administrators are also looking for a business model to sustain it. No mention is made of preservation though there is the assumption that the final product is deposited in a library, which by default will be responsible for long-term accessibility. (Kiondo 2004)

The Brazilian ETD initiative has as its goals (1) establishing standardized metadata; (2) implementing ETD-DL (digital library) to integrate various initiatives; and (3) distributing a software toolkit with implementation and training modules. Software and training has been extended to other Latin American countries, including Argentina, Colombia, and Uruguay, but there is no preservation strategy outlined. (Southwick and Pavani 2004)

KORDIC is the Korea Research and Development Center, which has had a project since 1998 to build a national digital library for ETDs. One of the few non-university based initiatives, it is in the Office of the Prime Minister. KORDIC's major function is to create a full-text database and a general distribution system accessible on the Internet. Most Korean universities request that their students submit ETDs, the result is a large number of diskettes stacked in the "warehouses of university libraries." (Lee 2001) A long-term preservation strategy is not addressed.

Documentation about the Indian ETD initiative describes the national goals to develop mechanisms and means for depositing, archiving, and accessing theses in India—to be a repository and an archive that draws on the Australian Digital Theses (ADT) model: collaborative, distributed, participatory digital library. (URS 2004)

In the ADT program, each university is responsible for maintaining an archival copy of their institution's theses on a local server. The institutions use the same database configuration, standards and metadata system to ensure compatibility (so implementing a preservation system like LOCKSS for ETDs would be very straightforward). The situation at the ADT program, like all others is an implicit archive, but lacking is the documented best preservation practices.

The emphasis in the United Kingdom has been almost entirely on dissemination. Preservation was hardly on the horizon anywhere except as a default library activity. (Andrew 2004) A recent announcement from the British Library declared development of a new national framework for the provision, preservation, and accession of theses. However, Electronic Theses Online is about access via the web to all theses electronically stored at the British Library along with information about holdings in other UK repositories. (Christensen 2005)

Since the late 1990s, the national library of Canada has contracted with UMI to reproduce theses on preservation quality microfiche. Library and Archives Canada (LAC) launched the

Theses Canada Portal in 2004 and provided free access to full text versions of theses and dissertations digitized without mentioning a digital preservation plan. (Theses Canada 2005)

The national library of Germany, Die Deutsche Bibliothek has an agreeable though fledgling preservation strategy that states the obvious: start new workflows using simple object categories like ETDs, i.e., monographic, finite, and stable without versioning problems. The conceptual views of the submission, archival, and dissemination information packages (i.e., SIP, AIP, DIP), are yet to be implemented. The German initiative also correctly states that the challenge is really organizational, not technical. (Liegmann 2003) Our ETD preservation projects using the LOCKSS software demonstrate this clearly and give every member of the NDLTD and every ETD initiative the opportunity to implement sound preservation strategies.

## 4. FUTURE REFINEMENTS

Future refinements for the MetaArchive Preservation Network must include strategies for enabling a technical infrastructure that supports an OAIS model (CCSDS 2002). This would involve the inclusion or expansion of the following components of the OAIS model: 1.) Common Services 2.) Data Management 3.) Archival Storage Layer and 4.) Administrative Tools. The further development of these areas would encourage replication for this type of preservation network by offering a framework around which many components of digital preservation could be managed offering an OAIS compliant structure based on modular additions to the network infrastructure.

**Common Services** in this context would include the incorporation of other modular tools that offer services that are beneficial both to the individual institution as well as to the preservation network. This would involve such services as file validation (JHOVE 2005), drive format compatibility and recovery (REISERFS 2005), and file format normalization and migration (DAITSS 2005).

**Data Management** is currently conducted using simple bit stream analysis with the LOCKSS software. Future enhancements would include integration with a Global Format Registry of file types as well as automated functions for file versioning and expanded capabilities for technical metadata. This would be useful for creating multiple versions of file types for use in future migration or emulation strategies.

**Archival Storage Layer** in its current state is very primitive and much more work will be necessary in order to perform automated tasks such as format migration or emulation. Currently the MetaArchive Network is based on collection level descriptions which do not capture enough metadata about individual objects to offer more than bit stream preservation. Utilizing tools such as JHOVE and DAITSS would require automated metadata creation from the ingest mechanism. This would be necessary in order to automate robust archival storage functions such as format migration or *on-the-fly* format conversion (Rosenthal et al. 2005).

**Administrative Tools** are currently built into LOCKSS for many of its original preservation goals and are based on the preservation of electronic journal archival units. Creating graphical administrative tools for file size limitations, archival unit creation, and for metadata manipulation will be necessary as more modular elements are added to the LOCKSS framework. This would include the OAI handling capability that is being developed in the LOCKSS daemon 1.10.

Obviously, the MetaArchive Network is one that is still in a highly developmental stage. Like many other digital preservation programs, ours is a first attempt at decentralized digital

preservation and will require further study and more importantly further practitioner-focused participation in order to fully realize its potential.

## 5. CONCLUSION AND CALL FOR PARTICIPATION

Now that the NDLTD has successfully fostered ETD initiatives worldwide, it is time for it to formally promulgate a preservation strategy. Among the ETD programs described here, there is as often as not the unspoken assumption that somebody else, generally "the library," is taking care of preservation. Often national efforts, such as the UK Digital Preservation Coalition; DINI (Deutsche Initiative für NetzwerkInformation), the German initiative for networked information; and NDIIPP at the U.S. Library of Congress are developing strategies for digital preservation but nothing specific is in place for ETDs.

The Algerian project is looking to initiatives such the "UNESCO ETDs Clearinghouse to encourage international communication and creation of spaces where it will be possible to share experiences, tools and ideas." (Bakelli and Benrahmoun 2003 p. 260) This may be a partial solution but does not address the need for preservation among the developed countries, which we have shown are also lacking in concerted digital preservation for their ETDs.

The NDLTD homepage says that it is "dedicated to … preservation … of electronic theses and dissertations." (http://www.ndltd.org) The NDLTD should follow the practical, working, and replicable approach to ETD preservation as demonstrated by the MetaArchive and ASERL projects. Therefore, we call on you to require it to follow through and put in place real preservation measures to ensure the long-term accessibility of ETDs. We advocate that the NDLTD form an international working group for the purpose of studying and proposing a preservation plan. We urge the NDLTD to form this group at this meeting and to set as a deadline ETD 2006 for implementation of the first phase of the plan.

Why not an Electronic Thesis and Dissertation International Preservation Network (ETD-IPN)?

## 6. REFERENCES


Andrew. T. (2004) Theses Alive! : An E-theses Management System for the UK. Retrieved July 15, 2005 from http://www.era.lib.ed.ac.uk/bitstream/1842/423/1/ASSIGN_thesesalive.pdf

Bakelli, Y., and Benrahmoun, S. (2003) Long-term preservation of ETDs in Algeria: Discussion through CERIST Deposit System. *Libri* 53(4):254-261. Retrieved July 25, 2005 from www.librijournal.org/pdf/2003-4pp254-261.pdf

Balile, D. (Oct. 23, 2003) Africa to Get Online Research Database. Retrieved July 15, 2005 from http://www.scidev.net/News/index.cfm?fuseaction-readNews&itemid=1068&language=1)

Christensen, L. (2005) Theses Unbound: Pioneering Electronic Access to UK Theses. (April 5, 2005) Retrieved July 25, 2005 from http://www.egovmonitor.com/node/412

Consultative Committee for Space Data Systems (CCSDS). (2002) Reference Model for an Open Archival Information System (OAIS), Blue Book, Issue 1, January 2002, ISO 14721:2003. Retrieved July 8, 2005 from http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.

DAITSS (2005) DAITSS Website. Retrieved July 20, 2005 from http://www.fcla.edu/digitalArchive/soft.htm.

DATAD (2004) Retrieved July 15, 2005 from www.aau.org/datad/database/ and http://www.aau.org/datad/cip/docs/DATAD_workshop_report.pdf and

Hedstrom, M., ed. (2003) It's About Time: Research Challenges in Digital Archiving and Long-term Preservation. Final Report of the Workshop on Research Challenges in Digital Archiving and Long-Term Preservation, Sponsored by the National Science Foundation and the Library of Congress, Washington, D.C. Retrieved July 8, 2005 from http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf.

Heratrix (2005) Heratrix Website. Retrieved July 20, 2005 from http://crawler.archive.org.

Internet2 (2005) Map of the I2 Abilene Network and Regional Optical Networks. Retrieved July 7, 2005 from http://www.internet2.edu/resources/AbileneMap.pdf.

JHOVE (2005) JHOVE Website. Retrieved July 20, 2005 from http://hul.harvard.edu/jhove.

Kiondo, E. (2004) Historical Practice in Managing Theses and Dissertations at African Universities and University Libraries. Retrieved July 15, 2005 from http://www.aau.org/datad/reports/2004workshop/kiondo.pdf

Lee, K. (2001) Construction of a Full-Text Database and Service System for Korean Electronic Theses and Dissertations. *Bulletin of the American Society for Information Science & Technology,* 27(3):21-27. Retrieved July 15, 2005 from http://www.asis.org/Bulletin/Mar-01/lee.html

Liegmann, H. (2003) Long-term Preservation of Electronic Theses and Dissertations. Retrieved July 25, 2005 from http://edoc.hu-berlin.de/conferences/etd2003/liegmann-hans/HTML/

Reich, V., & Rosenthal, D.S.H. (2001) LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine, 7(6)*. Retrieved July 1, 2005 from http://www.dlib.org/dlib/june01/reich/06reich.html.

RLG-OCLC. (2002) Trusted Digital Repositories: Attributes and Responsibilities. Mountain View, CA. Retrieved July 8, 2005 from http://www.rlg.org/en/pdfs/repositories.pdf.

Rosenthal, D., Lipkis, Robertson, T. & Morabito. (2005) Transparent Format Migration of Preserved Web Content. *D-Lib Magazine* 11(1): Retrieved July 7, 2005 from http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Southwick, S. B., and A. M. B. Pavani. (2004) Brazil's National Library of ETDs. Retrieved July 15, 2005 from http://www.uky.edu/ETD/ETD2004/abstract5.html

Theses Canada (2005) Retrieved July 25, 2005 from http://www.collectionscanada.ca/thesescanada/

Urs, S. (2004) Vidyanidhi Digital Library. Retrieved July 15, 2005 from www.aau.org/datad/reports/2004workshop/urs.pdf