



NDLTD/MetaArchive Preservation Strategy

Gail McMillan, Katherine Skinner
3rd ed. June 2010

Members of the NDLTD and the MetaArchive Cooperative have the opportunity to preserve their institutions' ETDs in a secure and private distributed preservation network. Because both organizations share the goal of enabling higher education institutions to provide long-term, open access to electronic theses and dissertations, the NDLTD is collaborating with the MetaArchive to co-sponsor a preservation strategy specifically for ETDs. This document outlines that strategy.

Current institutions engaged in the NDLTD/MetaArchive Preservation Strategy include Boston College, Emory University, Florida State University, Georgia Tech, Rice University, Virginia Tech. Incoming members include Auburn University, Pontifícia Universidade Católica-Rio de Janeiro, and University of Louisville.

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of electronic theses and dissertations (ETDs). It supports electronic publishing and open access to scholarship in order to enhance knowledge sharing worldwide. It provides resources for university administrators, librarians, faculty, students, and the general public. Topics include how to find, create, and preserve ETDs; how to set up an ETD program; legal and technical questions; and the latest news and research in the worldwide ETD community. <http://www.ndltd.org>

The MetaArchive Cooperative provides low-cost, high-impact preservation services to help ensure the long-term accessibility of the digital assets of universities, libraries, museums, and other cultural memory organizations. In addition to preserving members' digital content in a distributed digital preservation network, the Cooperative also offers consulting and education services to institutions that seek training in digital preservation planning, policy creation, and implementation, including setting up and running Private LOCKSS (<http://www.lockss.org>) Networks (PLN). <http://www.metaarchive.org>

NDLTD/MetaArchive Preservation Strategy

Table of Contents

1. Why adopt the NDLTD/MetaArchive preservation strategy?	3
2. MetaArchive Preservation Strategy	3
a. How the ETD preservation network works.	3
b. Access and the ETD Preservation Network	4
c. Intellectual Property Issues	4
d. Replacements available in the ETD Preservation Network	4
3. Organizing ETDs for Effective Collection Management and Preservation Readiness.....	4
a. Preserving Restricted and Withheld ETDs	5
b. Recommendations and Best Practices for Structuring New ETD Collections.....	5
c. Digitized Theses and Dissertations	6
4. Standards	6
a. File Formats	6
b. Metadata	6
5. Authors' Responsibilities	7
6. Institutional Workflow	7
7. Harvesting Frequency	8
8. How to Join the ETD Preservation Network.....	8
a. Join the NDLTD	8
b. Join the MetaArchive Cooperative	9
9. Documentation	10
10. Training: MetaArchive workshops.....	11
11. Staff/Personnel	12
12. Hardware	12
13. Software	12
14. Reports	12
Appendix A.....	13
1 st Draft: Recommendations and Best Practices for the ETD Preservation Network	13
Appendix B	14
ETD Collection Description "Template" for the MetaArchive Conspectus Database.....	14
Appendix C	18

NDLTD/MetaArchive Preservation Strategy

1. Why adopt the NDLTD/MetaArchive preservation strategy?

Essentially, all theses and dissertations created today are born-digital, and universities worldwide are accepting electronic theses and dissertations (ETDs) in addition to or in place of print versions. How we care for these new digital resources is important in light of possible catastrophic events such as fires and hurricanes, as well as the more prevalent hardware, software, and human failures that all institutions experience. We must be proactive in providing long-term digital preservation strategies to protect the important research and scholarship that comprises such an integral component of our institutional histories, or we run a high risk of losing it.

The MetaArchive Cooperative and the NDLTD joined forces in 2008 to offer preservation services for ETD collections by implementing a dark archive¹ using the technological approach referred to as a distributed digital preservation network (DDPN). In February 2010 a self-audit of the MetaArchive Cooperative determined that its distributed digital preservation network conforms to all 84 criteria specified by the Trusted Repositories Audit & Certification: Criteria & Checklist² (TRAC), and, therefore, operates according to the standards of a trustworthy digital repository. See MetaArchive Trusted Repository Audit,

http://www.metaarchive.org/sites/default/files/MetaArchive_TRAC_Checklist.pdf

Institutions may participate in the Cooperative by hosting a LOCKSS-based networked server. All collections will be ingested into the ETD Dark Archive by the MetaArchive system and copied, distributed, and stored on the secure servers of at least six NDLTD partner institutions in the MetaArchive Cooperative. These servers do not merely back up the ETDs, but provide a dynamic means of programmatically checking all files and providing replacement files when necessary.

2. MetaArchive Preservation Strategy

a. How the ETD preservation network works.

An adaptation of the Private LOCKSS Network (PLN), the Cooperative runs a DDPN, that consists of a group of at least six geographically dispersed servers (known as caches) that are networked through secure internet connections and a common software package: LOCKSS. These caches perform various functions: ingest content (e.g., ETDs) and store them; conduct polls to evaluate all caches' copies of the content to determine that all copies are intact and correct, repair or restore any content discovered through the polls that are no longer intact, periodically re-ingesting content from its original host in order to discover new or changed parts (without replacing the older version), and when necessary disseminating a copy of the content to an authorized recipient (usually to the institution from which it was originally harvested due to that file being damaged or corrupt due to natural disaster, technical failure or human errors).

¹ The purpose of a dark archive is to function as a repository for information that can be used as a failsafe during disaster recovery. In reference to data storage, it is an archive that cannot be accessed by any users.

(http://www.webopedia.com/TERM/D/dark_archive.html)

² Center for Research Libraries' *Metrics for Repository Assessment and Certification* <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying>

In summary, a PLN is a peer-to-peer system of geographically distributed servers (i.e., caches) running the open source software, LOCKSS, as its main application. These application processes are called LOCKSS daemons, i.e., computer programs running in the background without human intervention, that among other things poll each other about content, vote on the content to determine its integrity, and when necessary restore damaged content.

b. Access and the ETD Preservation Network

The NDLTD/MetaArchive Preservation Strategy separates preservation from user access. The MetaArchive preserves ETD collections in a dark archive, with access limited to specific NDLTD preservation partners. Content in the ETD Dark Archive is not accessible outside of MetaArchive's preservation routines, and is only available to the Cooperative's members for the distinct purpose of preservation and to restore to the originating institution its files when necessary.

User access to ETDs remains with the originating institutions or their designees. The ETD Dark Archive is solely a preservation network. The MetaArchive hosts a publicly available metadata repository that contains descriptions of each institution's collection(s). These metadata records are available to the public through the MetaArchive's Conspectus Database³ where there are links for the public to reach any collections available from the host university's access gateway.

c. Intellectual Property Issues

As a dark archive, the ETD preservation network is only accessible by specified MetaArchive servers and staff, and then only for purposes of preservation and replacing the originating institution's local files when necessary. All member institutions are individually responsible for ensuring that they have the appropriate level of copyright to provide distributed preservation for their ETDs, but they do not need to worry about access rights when participating in this preservation solution because it is strictly a dark archive.

The NDLTD and MetaArchive recommend that universities have their graduate student authors grant to their universities the non-exclusive license to archive and make accessible, under specified conditions, their theses and dissertations in whole or in part in all forms of media, now or hereafter known. Authors do not relinquish any rights and under these conditions they retain the copyright to their ETDs.

d. Replacements available in the ETD Preservation Network

Members may retrieve files when replacements are necessary for their local collections. The MetaArchive Program Manager is available to facilitate.

3. Organizing ETDs for Effective Collection Management and Preservation Readiness

Several participants in the MetaArchive Cooperative have contributed to an online book, *Guide to Distributed Digital Preservation* (<http://www.metaarchive.org/GDDP>). It includes in-depth and up-to-date information about best practices for organizing ETD collections with preservation in mind. In

³ <http://conspectus.metaarchive.org/archives/list>

addition, [Appendix A](#) of this document has a summary of recommended best practices. The synopsis below only addresses how to organize new ETD initiatives or ones in their early stages. For institutions that need to consider their existing ETD collections in the context of distributed preservation, information on triage is currently available in the paper Martin Halbert and Gail McMillan prepared for the 12th International Symposium on Electronic Theses and Dissertations, “Getting ETDs off the Calf-Path: Digital Preservation Readiness for Growing ETD Collections and Distributed Preservation Networks.”

<http://conferences.library.pitt.edu/ocs/viewabstract.php?id=733&cf=7>

All ETD collections must be web-accessible in order to be ingested and preserved in the ETD Dark Archive. Since harvesting for MetaArchive is based on LOCKSS, there must be an HTML permission-to-preserve statement that reads: “LOCKSS system has permission to collect, preserve, and serve this Archival Unit.” In LOCKSS terminology, this Manifest Page must exist the top-most level of the ETD directory in order to ingest the institution’s files into the ETD Dark Archive. See as an example the Virginia Tech ETDs LOCKSS Manifest Page at <http://scholar.lib.vt.edu/theses/lockss/manifest.html>

a. Preserving Restricted and Withheld ETDs

When access restrictions exist (for example, ETDs are available only to the host campus), each Preservation Node must add a specific preservation partners’ IP addresses to the web server's firewall configuration to allow secure access by only the specified servers (nodes) in the NDLTD preservation network.

b. Recommendations and Best Practices for Structuring New ETD Collections

To organize your institution’s ETDs most effectively for preservation harvesting, create a methodical structure such as a directory for each year’s ETDs. For institutions that approve ETDs that annually consume more than 20 GB, collections should be divided into logical units such as months or semesters. Smaller institutions whose annual ETD collections consume less than 20 GB will not benefit from creating these subdirectories.

Adopt the common and easy to decipher directory naming convention based on a timestamp⁴ assigned to each ETD: ddmmYYYY-ttttt. Preservation units (e.g., SIPs: Submission Ingest Packages or Archival Units in LOCKSS jargon) can, therefore, be based on year or year and month. These units are specifically for programmatic harvesting and not for human browsing. For the latter, you will most likely want to create browseable collections such as departments, authors, advisors, etc.

In LOCKSS terminology, these groupings are called “Archival Units.” They are fundamental to the computer scripts that enable collections to be ingested into the preservation network. Following these conventions for directory structure and file naming so that the base URL and year or month parameters will fit the ETD plugin examples available from Virginia Tech at <http://code.google.com/p/metaarchive/source/browse/#svn/branches/release/plugins/edu/vt/library>

The MetaArchive public LOCKSS network maintains the Cooperative’s plugins along with tools and support resources for plugin developers as well as a few scripts useful to LOCKSS

⁴ Timestamp here means a sequence of numbers denoting the date and time when an ETD was submitted.

cache administrators and administrators of web servers that 'deliver' content to LOCKSS caches in the network. See <http://code.google.com/p/metaarchive/>

c. Digitized Theses and Dissertations

Theses and dissertations that will be scanned should follow the above directory naming conventions. Let the name reflect the date of digitization rather than the date on which the ETD was originally approved or the degree was awarded. For example, is Ron Limoges's 1994 dissertation was scanned Oct. 2, 2007: <http://scholar.lib.vt.edu/theses/available/etd-10022007-144846/> It was harvested and preserved in the Archival Unit year=2007 along with the born-digital ETDs approved in 2007. This file naming schema also works well if an institution chooses to create archival units based on monthly designations.

4. Standards

a. File Formats

The MetaArchive distributed digital preservation network, like all LOCKSS networks, is format agnostic. That is, they can ingest any file formats into a preservation network. However, file formats will weather changes in versions and technology in a more consistent and community-supported manner if they are platform-independent, vendor-independent, non-proprietary, stable, and widely supported.

Migration is a preservation strategy that transforms a file to create a new version of that file in a different format, where the new format is compatible with contemporary software and hardware. Ideally, migration is accomplished with as little loss of content, formatting, and functionality as possible, but the amount of information loss will vary depending on the original formats and content types involved.⁵

The MetaArchive Cooperative provides bit-level preservation for all files ingested into its preservation network. In the future, MetaArchive may develop and guarantee migration strategies for all files in standard, robust formats (see below) that are ingested into the ETD Dark Archive as formats become obsolete. It may also make its best effort to migrate non-standard formats that have an established community of practice. MetaArchive also welcomes files that the contributing institution will migrate itself (e.g., a database that a student developed in an unsupported format).

The file formats that the NDLTD recommends are also recommended by the MetaArchive as those most likely to migrate readily to subsequent standards. These are text formats: PDF; image formats: TIF, JPG, GIF, and PNG; video formats: MPG, MOV, QT; and audio formats: WAV, MPG, and MP3.

b. Metadata

The Conspectus Database describes each collection. This collection-level metadata serves to both inform the public broadly about the collections held in the ETD Dark Archive and to facilitate MetaArchive network management, harvesting, maintenance, and recovery routines.

⁵ *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group* (May 2005) <http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>

The descriptive elements adopted by the MetaArchive provide basic information about each collection and its context as well as administrative information about the format(s) of the content. The MetaArchive adapted robust and well-respected existing schemas to create the MetaArchive Conspectus Schema.⁶ While the Conspectus Database has brief summaries of each piece of metadata, the Schema thoroughly defines each element.

See Appendix B, which includes information for the MetaArchive's Conspectus Database about a generic ETD collection. Also, consult live examples of the ETD collections <http://conspectus.metaarchive.org/archives/1>

5. Authors' Responsibilities

ETD authors have considerable influence on long-term access and preservation. They should be educated and trained to produce works that are easy for readers to navigate and read, both now and in the future. These authors often need to be informed to consider preservation issues that will enable long-term access to their ETDs. For example, they should always include the highest resolution of an object (e.g., image or audio file) rather than a version that is suitable for today's devices because the technology will improve. They should include a version using a well-accepted international standard. Acceptable file formats will stand up to changes in versions and technology better if they are platform-independent, vendor-independent, non-proprietary, stable, and widely supported. See 4.a. above.

Intellectual property law is a national consideration and teaching ETD authors about their rights and responsibilities is often not given enough attention. While they frequently want to know if they can use copyrighted texts, tables, charts, illustrations, surveys, etc., they are often unaware of their rights as creators of new works. Both sides of the copyright question should be addressed in ETD training situations.

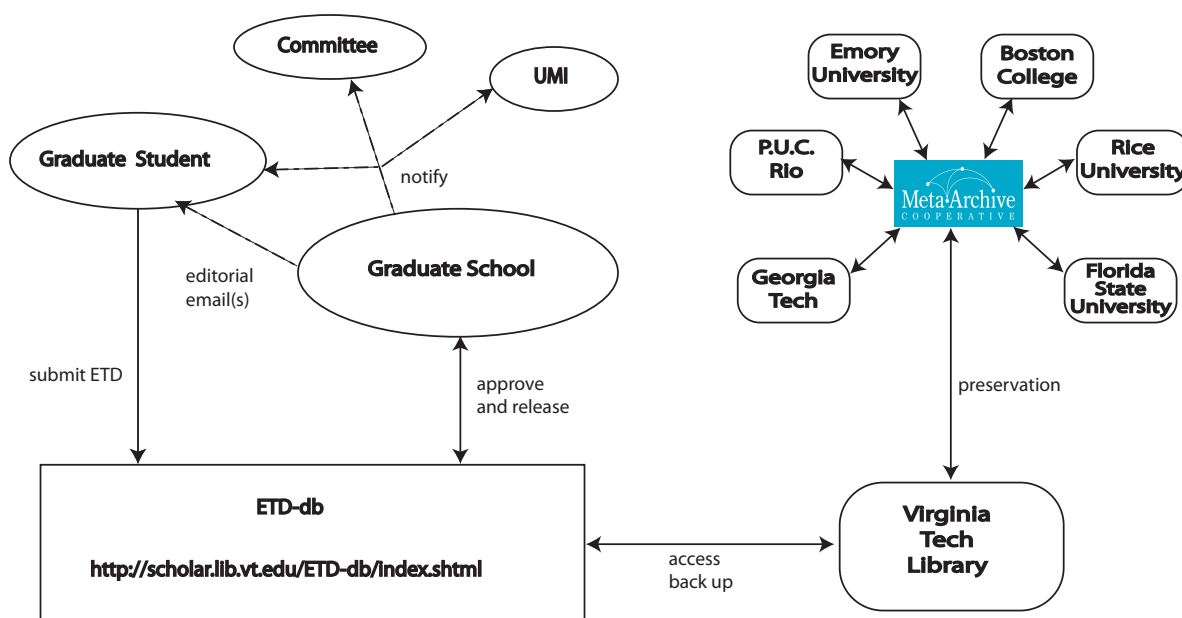
See 2.c. above, Intellectual Property Issues for the NDLTD and MetaArchive's recommendations regarding graduate student authors granting to their universities a non-exclusive license to archive and make accessible, under specified conditions, their theses and dissertations. The authors still retain ownership rights (i.e., copyright) to their ETDs.

6. Institutional Workflow

Each university determines its local policies and procedures, developing a work flow from submissions to approval, accessibility, and preservation that best fits its particular needs. Here is one example incorporating ETD-db software and employing the MetaArchive's preservation network.

⁶ http://metaarchive.org/sites/default/files/conspectus_md_2005.html

ETD-db Flow Diagram with Preservation Network



7. Harvesting Frequency

ETDs are submitted to meet periodic institutional deadlines. Therefore, each institution should analyze the rate of change of its ETD collection and determine the most appropriate frequency for preservation harvesting into the ETD Dark Archive. MetaArchive recommends annual harvesting so that the Archival Unit is static and complete (i.e., all of the ETDs submitted in a single year) but only if the AU will be smaller than ten gigabytes. Annual ETD collections that are larger than 20 GB should be divided into logical units and harvested more frequently, probably twice a year. See “Organizing ETDs for Effective Collection Management” in 3 above. Routine backups should continue to be a standard operating procedure.

8. How to Join the ETD Preservation Network

A five percent discount is available to institutions that join the NDLTD and the MetaArchive Cooperative together. Membership is required in both organizations to participate in the ETD preservation network.

a. Join the NDLTD

The Networked Digital Library of Theses and Dissertations is an international, non-profit organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of ETDs and digital analogs to the traditional paper-based theses and dissertations. Since its inception in 1997, the 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.

An elected Board of Directors guides the NDLTD and works with representatives from member institutions on various committees to further the aims of the organization, including digital preservation. At its Dec. 10, 2008 meeting the Directors enthusiastically endorsed establishing a distributed preservation network for ETDs through the MetaArchive Cooperative.

Membership in the NDLTD⁷ is required of all who wish to take part in the ETD Dark Archive. Annual institutional membership fees range from \$100-\$300, with a sliding scale for university systems and consortia, and lower fees for institutions in developing countries (per the 2003 United Nations Human Development Report).

b. Join the MetaArchive Cooperative

The MetaArchive Cooperative is a service and preservation association designed for the mutual benefit of its members. It consists of academic and other nonprofit institutions that share a common goal of preserving digital scholarly and cultural resources for the future. The core mission of the MetaArchive Cooperative is to support, promote, and extend the practice of distributed digital preservation; to serve as a catalyst and guide for other networks that seek to implement the distributed digital preservation methods it has developed; and to educate organizations about distributed digital preservation.

The MetaArchive Cooperative provides low-cost, high-impact preservation services to help ensure the long-term accessibility of the digital assets of cultural memory organizations, such as universities and their ETDs. MetaArchive was formed in 2004 out of increasing concern for the digital items that define our culture and history that could be lost due to natural disaster, human error, or sheer neglect. The MetaArchive Cooperative functions as a community initiative. The newly launched NDLTD distributed digital preservation network will cooperatively preserve ETD collections, not by outsourcing, but by actively participating in the preservation of our institutional research and heritage.

Membership in the MetaArchive Cooperative is required of all who wish to take part in the ETD preservation network. Two membership categories are available based on the extent of institutional participation in the network. Each membership level has an allotment of space for preservation of each ETD collection. For details about membership as well as mission, goals, and principles, contact⁸ Martin Halbert, President of MetaArchive Services Group and Director of the Digital Programs and Systems at Emory University, or the MetaArchive Program Manager, Katherine Skinner. Potential members may consult the Cooperative's Charter⁹ and its Membership Agreement¹⁰ that includes the contract with members that frames the relationship.

MetaArchive Membership Levels and Activities

Preservation Members

In addition to having their ETD collections harvested and cached by the ETD Dark Archive, Preservation Members also harvest and cache web accessible ETDs for other NDLTD

⁷ <http://www.ndltd.org/join/>

⁸ <http://www.metaarchive.org/contactus/>

⁹ http://www.metaarchive.org/public/resources/charter_member/MetaArchive_Charter_2010.pdf

¹⁰ http://www.metaarchive.org/public/resources/charter_member/Membership_Agreement_2010.pdf

institutions participating in the ETD Dark Archive. This level of membership requires that an institution maintain a server or “network node” for the Cooperative that meets specific technical requirements.¹¹ They also administer systems and monitor harvesting and caching procedures, which are part of the LOCKSS programmatic routines. Once a Preservation Member establishes its operational node, background tasks require minimal amounts of staff time.

The institution fee for Preservation Members is \$1,000 per year for an initial three-year commitment. Included in the fee is 20 GB of space for each Preservation Member’s collection in the ETD Dark Archive.

Sustaining Members

The MetaArchive’s Sustaining Members have the widest array of responsibilities, but they also have the greatest influence over the development of the MetaArchive Cooperative through their representation on the Steering Committee. Along with the responsibilities of Preservation Members, Sustaining Members also test and develop hardware, software, networking, and transmission standards, and they research and deploy the work of the Cooperative, contributing staff and resources.

The institutional fee for joining the MetaArchive Cooperative as a Sustaining Member is \$5,000 per year. However, institutions that pay for a three-year membership at the time that they sign their membership agreement earn a \$1,000/year discount, thus paying a total of \$12,000 for a three-year institutional membership instead of \$15,000 paid over three years. Included in the membership fee is access to 40 GB of space for each Sustaining Member’s collection in the ETD Dark Archive.

LOCKSS¹² Alliance

Sustaining and Preservation Members of the MetaArchive Cooperative are required to maintain membership in the LOCKSS Alliance. Institutional membership fees vary and are based on country and size of institution.¹³

9. Documentation

A wealth of information is available in the MetaArchive’s online book, *A Guide to Distributed Digital Preservation*, and there are a wealth of online resources documenting the MetaArchive Cooperative and its activities at its Web site, <http://www.metaarchive.org>, including:

Cooperative Charter: provides the mission, goals, and organizing principles of the MetaArchive Cooperative. It outlines membership levels and details the roles and responsibilities in this collaborative model.

http://www.metaarchive.org/public/resources/charter_member/MetaArchive_Charter_2010.pdf

Membership Agreement: defines the terms of the contract made between members of the MetaArchive Cooperative. It provides a framework for these relationships while remaining flexible enough to address the evolving needs of the Cooperative.

¹¹ <http://www.metaarchive.org/pdfs/AppendixA0208.pdf>

¹² <http://www.lockss.org>

¹³ [http://www.lockss.org/lockss/LOCKSS Alliance](http://www.lockss.org/lockss/LOCKSS_Alliance)

http://www.metaarchive.org/public/resources/charter_member/Membership_Agreement_2010.pdf

Technical Specifications: details the network specifications and technical infrastructure required of Preservation and Sustaining Members.

http://www.metaarchive.org/sites/default/files/Appendix_A_Technical_Specifications.pdf

A Guide to Distributed Digital Preservation: describes successful, low-cost collaborative strategies and provides specific new models that can help higher education institutions work together for their mutual benefit. Use it to gain both a philosophical and a practical understanding of the emerging field of distributed digital preservation, including how to establish or join a network.

<http://www.metaarchive.org/gddp>

Management Plan: includes detailed information regarding the ownership structure of the Cooperative, its internal and external human resources, and the human resources the Cooperative anticipates it will need in the future to sustain its activities.

http://www.metaarchive.org/public/resources/ndiipp_docs/NDIIPP_Management_Plan.pdf

Extension Harvest Plan: charts the network development and harvest plans for the overall MetaArchive network, including building on the Southern Digital Culture archive and founding two new archives.

http://www.metaarchive.org/sites/default/files/Extension-Harvest-Plan-FINAL_0208.pdf

Outreach Program Implementation Plan: describes goals and activities for the MetaArchive Cooperative's community building through March 2010.

http://www.metaarchive.org/sites/default/files/OutreachProgramImplementationPlan_final_0108.pdf

Data and Node Recovery: provides an in-depth guide to LOCKSS data and node recovery activities that have been thoroughly tested and that may be needed by PLNs in case of node failures.

http://www.metaarchive.org/sites/default/files/Network_Test_0208_Disaster_Recovery_FINAL.pdf

10. Training: MetaArchive workshops

The second MetaArchive workshop specifically designed for the NDLTD and the ETD Dark Archive was held in conjunction with the 12th International Symposium on Electronic Theses and Dissertations at the University of Pittsburgh on June 10, 2009.¹⁴ These workshops have become annual half-day preconference workshops at the ETD symposia, which provide participants with information regarding the development of the NDLTD/MetaArchive partnership and why distributed digital preservation is an important safeguard for ETD collections, as well as the explaining the Cooperatives services and institutional members' roles.

The MetaArchive also conducts workshops periodically so that a broad range of institutions may learn more about the technical logistics and operational considerations of hosting or participating in a Private LOCKSS Network (PLN) for distributed preservation. **The next technical workshop will be held in conjunction with the MetaArchive Steering Committee meeting in Boston in October 2010. IS THIS TRUE?** Typically the all-day workshops provide extensive information and training for institutions seeking to join the Cooperative's LOCKSS-based distributed digital preservation network or to build their own DDPN based on the MetaArchive model.

¹⁴ <http://www.library.pitt.edu/etd2009/pdfs/ETDPreservationWorkshop200906.pdf>

11. Staff/Personnel

Institutional staffing for participation in the ETD Preservation Network will vary with the category of MetaArchive membership selected. Sustaining and Preservation Members will require some staffing resources to set up, manage, and monitor the MetaArchive nodes that they will run. Sustaining Members devote additional time to the administration of the Cooperative. For more information, please contact the MetaArchive Cooperative's Program Manager. ([See 15 below](#))

12. Hardware

Preservation and Sustaining Members must meet the MetaArchive's hardware requirements and technical specifications for running a server node that are detailed and kept current at http://www.metaarchive.org/sites/default/files/Appendix_A_Technical_Specifications.pdf

13. Software

Preservation and Sustaining Members employ the LOCKSS software and additional components developed by MetaArchive to establish, run, and monitor their networked nodes. These are open source and available online without charge.

14. Reports

Members of the MetaArchive Cooperative have access to many reports that are generated programmatically and available online through the Conspectus Database and Cache Manager.

15. **Contact MetaArchive Program Manager** for specific information about institutional memberships and with questions about categories and levels of participation in the Cooperative.

Katherine Skinner, PhD
Executive Director, Educopia Institute
Program Manager, MetaArchive Cooperative
404 783 2534; katherine.skinner@metaarchive.org

Appendix A

1st Draft: Recommendations and Best Practices for the ETD Preservation Network

1. Effectively Organizing an ETD Collection

- a. Create a broad-based logical collection structure.
- b. Define standardized naming conventions for files and directory structures from the beginning.
- c. Adopt a uniform, regular, and easy to decipher naming convention
 - i. Accumulation periodicity, such as a directory for each year's ETDs
 - ii. Timestamps, etd-mmddyyyy-tttttt, when students submit their ETDs
 - iii. If annual collections are >20 GB, divide directories into monthly subunits.
- d. Scanned (versus born-digital) theses and dissertations follow the same directory naming convention based on the timestamp for the digitization date.

2. Triage for Legacy Collections

- a. A general strategy of remediation: recognize and put boundaries around an irregular collection that requires special measures for data management.
- b. Cease adding to unorganized collection, thus creating a static collection with a finite number of files, and begin to implement new best practices based on the above logic.
- c. Create a (*virtual and artificial*) collection with just one archival unit.

3. Live versus Static Media

- a. Store ETDs on live, spinning discs, not on CDs or other static storage devices.
- b. Eliminates having to find those discs, load them onto spinning discs, rectify errors and failed media.

4. Metadata Discipline

- a. Preserving ETDs means not only preserving the files that comprise these works.
 - i. Preserve the metadata that describes each work.
 - ii. Preserve the scripts or database structure that manages them.
- b. Associate sufficient metadata with digital assets that they can usefully be accessed and managed by subsequent generations of staff and users.
 - i. ETD-MS serves as the standard for theses and dissertations.
 - ii. It is standard practice for libraries to catalog theses and dissertations
 1. ETDs can have ready-made metadata derived from MARC, or
 2. Metadata can be derived from the descriptive elements authors enter when they submit their ETDs (and used to generate MARC)

5. File Formats

- a. Include the highest resolution of any object (e.g., image or audio file) rather than a version that is suitable for today's devices because the technology will improve.
- b. Include a version using a well-accepted international standard.
- c. Acceptable file formats will stand up to changes in versions and technology better if they are platform-independent, vendor-independent, non-proprietary, stable, and widely supported.
- d. Text formats: PDF; image formats: TIF, JPG, GIF, and PNG; video formats: MPG, MOV, QT; and audio formats: WAV, MPG, and MP3

Appendix B
ETD Collection Description "Template" for the MetaArchive Conspectus Database

Public	Required	Sections and Elements	Brief Summaries of Sections and Element Descriptions	Describe your ETD Collection, for example:
		Collection Description Data Creator/Editor	Complete this form to give an overview of the collection to be preserved and provide elements (metadata) essential for harvesting. Conspectus Schema has more information: http://metaarchive.org/pdfs/conspectus_md_2005.html	
		Descriptive Data	Details explaining the digital collection to be preserved	
P	R	Collection Title	What is the formal name of this group of materials?	<i>ETDs@[name of university]</i>
P		Alternative Title	Other names for this collection	<i>Electronic Theses and Dissertations at [name of university]</i>
P	R	Description	Explain or define this group of materials	<i>[This is a growing collection of electronic theses and dissertations, documents similar to their paper predecessors that explain the research of [name of institution]'s graduate students. ETDs have been [choose one: optional or required] since [date]. [In addition to born-digital ETDs, this collection also contains digitized theses and dissertations.]</i>
P	R	Subjects	Describe the collection using terms from a thesaurus or controlled vocabulary (e.g., Library of Congress Subject Headings).	<i>[Name of institution] -- Research. [Name of institution] -- Graduate Students.</i>
		URIs	Uniform Resource Identifier (URI)--usually a locator (URL) or a name (URN).	
	R	Collection URI	URI associated with this group of digital materials	<i>http://scholar.lib.vt.edu/theses/</i>
		Institution Identifier	University assigned control number or name for the digital collection	
P		Is available via	URL where the digital collection is available to users.	<i>http://scholar.lib.vt.edu/theses/</i>

Public	Required	Sections and Elements	Brief Summaries of Sections and Element Descriptions	Describe your ETD Collection, for example:
		Coverage	Describe the collection in space and time	
		Spatial Coverage	Geographical location (place or areas) associated with the contents of the digital collection	
		Temporal Coverage	Time period associated with the contents of the digital collection	
P		Accumulation Date Range	Span of dates during which the collection was assembled	1996-
P		Contents Date Range	Dates of creation of the digital collection	1995-
		Accrual Information	Information relating to the growth (accumulation) of the collection	
P		Accrual Periodicity	Frequency with which items are added to a collection	
		Select only one	Choices: No Longer Adding; Daily; Weekly; Monthly; Quarterly; Yearly; Occasionally	YEARLY
P		Accrual Policy	Approach adopted to add items to the collection or a statement about the anticipated growth of the collection	<i>[Files are added as the Graduate School approves ETDs and as bound volumes are scanned for online access.]</i>
		Data Description	Formatting, size, and language information associated with the collection	
	R	Format Characteristics	Physical or digital characteristics of the files in the collection	<i>[Audio: wav; Image: gif; Image: jpg; text: pdf; text: html; text: xml; ...]</i>
P		Language	Language of the content of the items in the collection	<i>[English, French, Spanish...]</i>
P		Type	Nature or genre of the content of in the collection	
		Select more than one.	Computer Animation; Complex or Learning Objects; Databases; Datasets; Events; Interactive Resources; Moving Images; Physical Object; Services; Software; Sound; Still Images; Text	<i>[Databases, datasets, moving images, software, sound, still images, text]</i>
P	R	Extent Bytes and Collection Size	Size or duration of the entire digital collection expressed in [radio buttons for B, Kb, Mb, Gb, Tb) as of yymmdd	<i>[100931731456]</i>

Public	Required	Sections and Elements	Brief Summaries of Sections and Element Descriptions	Describe your ETD Collection, for example:
		Rights and Ownership	Information about intellectual property (copyright)	
P	R	Creator	Originator of the content in the digital collection	[Name of university]
P	R	Publisher	Entity responsible for making the resource available	[Name of university]
P	R	Rights	Statement about who owns the copyright	[Available for research, teaching, and private study... ETDs are not in the public domain and copyright is largely held by... print ... provide proper attribution ... http://]
P	R	Access Rights	Statement of restrictions placed on the collection, including allowed users, charges, etc.	
		Select more than one.	Unrestricted; restricted	RESTRICTED
P		Custodial History	Statement about changes in ownership and custody of the digital collection that are significant for its authenticity, integrity and interpretation; include provenance	[ETDs have lived on the DLA server at scholar.lib.vt.edu/theses/ since the initiative began in 1995.]
		Manifestation	Indicate role of the files in collection--reformatting quality attribute.	
		Select more than one.	Access; Preservation; Replacement	[Preservation; Replacement]
		Related Resources	Data concerning the use of and references for the collection	
		Associated Publications	Bibliographic citation (and URL) of publications based on use, study, or analysis of the collection	
P		Subcollection	Name of collection (and URL) that is contained within this collection	Browse by author: http://
P		Supercollection	Name of collection (and URL) that contains this digital collection	[Digital Library and Archives: http://scholar.lib.vt.edu]
		Catalog or Description	Finding aid or other publication that provides intellectual access to this digital collection.	MetaArchive Conspectus Database
		Cataloged Status	Detailed description of the cataloging or other metadata available for items in the collection.	[The ETD collection is not cataloged, but each ETD has a record in the local catalog, http://addison.vt.edu and in WorldCat.]
		Select only one	Catalogued: [yes, no, partially]	[Not Catalogued]

		Associated Collection	Name(s) of collection(s) associated by content or provenance	
Public	Required	Sections and Elements	Brief Summaries of Sections and Element Descriptions	Describe your ETD Collection, for example:
		Harvesting Information	Information about the web crawl that will gather the files for archiving	
	R	Harvest Procedure	Select method for gathering files into the archives.	
	R	Select only one	Webcrawl; OAI Harvest	<i>Webcrawl/</i>
	R	Plugin Identifier	Provide the URI/URL for the plugin.	<i>[edu.vt.library.theses]</i>
		Extra Parameters	Create entries for each Archival Unit. Base_URL is automatically included.	<i>[year=1997, ... year=2008]</i>
	R	LOCKSS Manifest Page	URL of the permission statement that enables harvesting	<i>[http://scholar.lib.vt.edu/theses/lockss/manifest.html]</i>
		OAI provider	Specifies the URL of the Open Archives Initiative data provider	
	R	Risk Rank	Designate the degree to which the collection is in jeopardy	
		Select only one	Extreme Risk: No one is responsible for preservation. No other copies are preserved beyond the copy under consideration. No regular backups or data migration	<i>[Extreme Risk]</i>
			Significant Risk: Responsibility under discussion.	
			High Risk: Only one backup on CD-ROM.	
			Moderate Risk: Danger that collection backups might be lost.	
			Low Risk: Copies are backed up regularly with a long term maintenance plan in some other trusted digital archive	
		Risk Factors	Describe the reason this collection is endangered.	<i>[ETDs are stored at only one campus site. Born-digital ETDs are only available electronically. Scanned works may have a bound paper copy in Special Collections.]</i>

Appendix C

ETD Preservation in the MetaArchive Cooperative

