NDLTD Preservation Strategy with the MetaArchive Cooperative

Gail McMillan, Katherine Skinner
May 19, 2008

The NDLTD and the MetaArchive Cooperative share the goal of helping higher education institutions provide long-term open access to electronic theses and dissertations. This document strives to outline and provide enough details concerning how NDLTD members can achieve this goal by becoming part of a new ETD preservation network.

Table of Contents

1. Why Participate
2. How to Join
3. Hardware
4. Software
5. Access
6. Intellectual Property
7. Organizing Strategies for ETD Collections
8. Standards
9. Harvest Frequency
10. Institutional Workflow
11. ETD Authors’ Responsibilities
12. Personnel
13. Training Opportunities
14. Documentation and Reports
15. Retrieving from the ETD Preservation Network
16. Contact Program Manager

Important Links to Web Documents

Networked Digital Library of Theses and Dissertations
http://www.ndltd.org/
MetaArchive Cooperative
http://www.metaarchive.org
NDLTD Preservation Strategy with the MetaArchive Cooperative
http://scholar.lib.vt.edu/theses/preservation/NDTLDPlan.pdf
MetaArchive Conspectus Database
http://www.metaarchive.org/conspectus/
MetaArchive Conspectus Schema
The NDLTD and the MetaArchive Cooperative share the goal of helping higher education institutions provide long-term open access to electronic theses and dissertations. This document strives to outline and provide enough details concerning how NDLTD members can achieve this goal by becoming part of a new ETD preservation network.

1. Why Participate in an NDLTD MetaArchive Preservation Network

More than 93% of the world’s information today originates as digital files, not print documents. Essentially all theses and dissertations created today are born-digital even if they are reviewed and read as printed works. Increasingly, 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 encounter. 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 begin offering preservation services to preserve ETD collections by implementing an ETD archive using the technological approach called distributed digital preservation network. Participants in this new archive may simply submit their collections for ingest into the network or they may participate in the Cooperative by hosting a LOCKSS-based networked server. All collections will be ingested into the ETD Archive by the MetaArchive system and copied, distributed, and stored on secure servers at multiple NDLTD partner institutions. These servers do not merely back up the ETDs, but provide a dynamic means of continually programmatically checking all files and providing replacement files when necessary.

2. How to Join: Memberships Required in NDLTD and MetaArchive Cooperative [$300-$5,300]

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 electronic analogues to the traditional paper-based theses and dissertations. Since its inception in 1996, 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 Jan. 21, 2008, meeting the Directors enthusiastically endorsed establishing a distributed preservation network for ETDs through the MetaArchive Cooperative.

Membership in the NDLTD\(^1\) is required of all who wish to take part in the ETD preservation network. 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).

**MetaArchive Cooperative**  
http://www.metaarchive.org/

The MetaArchive Cooperative is a service and preservation association designed for the mutual benefit of all 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 universities such as 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. Three 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, contact the Program Manager\(^2\) and/or consult the MetaArchive Cooperative Charter\(^3\) and Membership Agreement\(^4\) that includes the contract with members that frames the relationships.

**Contributing Members**

Institutions that join the MetaArchive Cooperative as Contributing Members contract for services only and thus assume no preservation responsibilities beyond preparing their own ETDs for harvesting. These institutional members do not have any technical obligations, nor do they have an active role in the operation of the Cooperative.

---

2. kskinne@emory.edu  
4. [http://metaarchive.org/pdfs/MetaArchiveMembershipAgreement.pdf](http://metaarchive.org/pdfs/MetaArchiveMembershipAgreement.pdf)  
5. [http://metaarchive.org/resources.html](http://metaarchive.org/resources.html)
The introductory membership fee for Contributing Members joining the MetaArchive Cooperative is $200 US per year [as of April 15, 2008] for a three-year term, which includes guidance from the Cooperative to ensure that the institution’s ETD collection(s) are harvestable, the writing of one fully functioning plug-in (a required ingest element), and five gigabytes of storage.

**Preservation Members**

In addition to having their ETD collections harvested and cached by the ETD Archive, Preservation-level Members also harvest and cache those of other NDLTD institutions. This level of membership requires that an institution maintain a server or “network node” for the Cooperative that meets specific technical requirements. 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, unobtrusive tasks require minimal amounts of work time.

The institution fee for Preservation Members is $1,000 per year for an initial three-year commitment. Included in the fee is access to 20 GB of space for each Preservation Member’s collection in the ETD 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. Included in the membership fee is access to 40 GB of space for each Sustaining Member’s collection in the ETD Archive.

**LOCKSS 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.

3. **Hardware**

Contributing Members do not have any hardware requirements, as they do not run servers for the MetaArchive. Preservation and Sustaining Members must meet the MetaArchive’s technical specifications.

---

8 [http://www.lockss.org](http://www.lockss.org)
9 [http://www.lockss.org/lockss/LOCKSS_Alliance](http://www.lockss.org/lockss/LOCKSS_Alliance)
11 [http://metaarchive.org/resources](http://metaarchive.org/resources)
4. Software

Contributing Members do not have any preservation software requirements. Preservation and Sustaining Members employ the LOCKSS software\(^\text{12}\) and additional components developed by MetaArchive\(^\text{13}\) to establish, run, and monitor their networked nodes.

5. Access

The NDLTD MetaArchive preservation strategy currently disconnects preservation from public web access. At this time the MetaArchives preserves ETD collections in a dark archive, with access limited to specific ETD archive partners. Content in the ETD 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 replace the originating institution's local files when necessary.

Access gateways for the public are individually hosted and monitored by each university (separate from the preservation strategy). However, the MetaArchive hosts a publicly available metadata repository that contains descriptions of each university’s ETD collection. These metadata records are available to the in the MetaArchive Conspectus Database\(^\text{14}\) where there are also links for the public to reach any collection available from the host university’s access gateway.

Future development of the MetaArchive Cooperative may include adopting the LOCKSS feature that enables public access from the preservation network. This would be each university’s decision to enable access and to share unrestricted ETDs with the public from the preservation network if the host university’s access gateway becomes inoperable.

Metadata

Each institution member must describe its ETD collection in the MetaArchive Conspectus Database. The resulting descriptions serve to both inform the public broadly about the collections held in the ETD Archive and facilitate the MetaArchive network management, harvesting, maintenance, and recovery routines.

The descriptive records developed 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\(^\text{15}\). While the Conspectus has brief summaries of each piece of metadata, the Schema thoroughly defines each element.

With this packet there is a Conspectus Spreadsheet for ETDs that has ETD-specific information for creating collection-level metadata.

\(^{12}\) [http://www.lockss.org/lockss/Installing_LOCKSS](http://www.lockss.org/lockss/Installing_LOCKSS)
\(^{13}\) [http://www.metaarchive.org/adaptedMetaArchiveSoftware](http://www.metaarchive.org/adaptedMetaArchiveSoftware)
\(^{14}\) [http://www.metaarchive.org/conspectus/](http://www.metaarchive.org/conspectus/)
6. 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 must ensure that they have the appropriate level of copyright to provide distributed preservation for their ETDs, but do not need to worry about access rights when participating in this preservation solution.

The NDTLD 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. The authors still retain the copyright to their ETDs.

7. Organizing ETDs for Effective Collection Management

The MetaArchive wiki above provides in-depth and up-to-date information about best practices for organizing ETD collections with preservation considerations in mind. The synopsis below only addresses how to organize new ETD initiatives or ones in their early stages. For institutions that now must consider their existing ETD collections in the context of distributed preservation, information will be available that specifically addresses ETD_db, DSpace, Fedora, and CONTENTdm.16

All ETD collections must be web-accessible in order to be ingested and preserved in the ETD Archive. Since harvesting for MetaArchive preservation is based on LOCKSS, there must be a standard HTML permission-to-preserve statement (i.e., Manifest Page) at the top-most level of the ETD directory in order to ingest the institution’s files into the MetaArchive.17

Preserving Restricted and Withheld ETDs

When access restrictions exist (for example, only the host campus has access permissions), a list of specific preservation partners’ IP addresses must be added to the web server's firewall configuration to allow secure harvesting by only the specific servers (nodes) on the NDLTD preservation network.

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 larger institutions that approve hundreds of ETDs each year, annual directories should be further subdivided into logical subunits such as months. Smaller institutions that approve 100 or fewer ETDs per year will not benefit from creating these subdirectories.

Adopt the common and easy to decipher naming convention year/month. Use 2008/01 to designate January 2008; do not use 2008January. 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.

---

16 https://www.metaarchive.org/metawiki/
17 See an example at http://scholar.lib.vt.edu/theses/lockss/manifest.html
In LOCKSS jargon, these groupings are called “Archival Units.” They are fundamental to the computer scripts that ingest collections into the preservation network. Contributing Members must follow these conventions for directory structure and file naming so that the base URL and year parameters will fit the plugin master/template. This will enable the MetaArchive Cooperative to provide Contributing Members with plugins. Otherwise, each university must create a plugin specific to the structure of its ETD collection.

**Digitized Theses and Dissertations**

Theses and dissertations that will be scanned should also follow the above naming convention. The subdirectory name may reflect the date of digitization rather than the date on which the ETD was originally approved or the degree was awarded. For example, http://scholar.lib.vt.edu/theses/available/etd-10022007-144846/ is Ron Limoges’s 1994 dissertation at Virginia Tech that was scanned Oct. 2, 2007. It was harvested and preserved with the born-digital ETDs approved in 2007.

**8. Standards**

The MetaArchive distributed digital preservation network is format agnostic and ingests all file formats into its preservation archive. 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.18

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

The file formats recommended by the NDTLD are also recommended by the MetaArchive as most likely to be migratable 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.

**9. 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 and caching into the ETD Archive.

---

10. Institutional Workflow

Each university determines its local policies and procedures, developing a work flow from submissions to approval and accessibility that best fits its particular needs. Routine backups should continue to be standard operating procedures at each university. Institutions will most likely want to create collections that can be browsed by departments, authors, advisors, and perhaps degree and year.

Each institution must organize its ETD collection for effective harvesting and ingest, i.e., efficient preservation. See “Organizing ETDs for Effective Collection Management” above.

11. ETD Authors’ Responsibilities

Authors also have opportunities to enable long-term access and preservation. Student authors should be educated and trained to produce works that are easy for readers to find, navigate, and read, both now and in the future. Students often need to be informed to be thoughtful and consider issues of preservation that will enable long-term access to their ETDs. For example, they should always include the highest resolution of an object 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.

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 6 above, Intellectual Property Issues for the NDTLD and MetaArchive’s recommendations regarding having graduate student authors grant to their universities a non-exclusive license to archive and make accessible, under specified conditions, their theses and dissertations. The authors still retain all ownership rights to the copyright of their ETDs.

12. Personnel

Institutional staffing for the ETD Archive will vary with the category of MetaArchive membership selected. Contributing Members will not require staffing beyond organizing their ETD collections for access and harvesting by the preservation network. Sustaining and Preservation Members will require some staffing resources to set up, manage, and monitor the MetaArchive nodes that they will run. For more information, please contact the MetaArchive Cooperative’s program manager.  

13. Training: MetaArchive workshops

The first MetaArchive workshop specifically designed for the NDTLD and the ETD Archive is being held in conjunction with the 11th International Symposium on Electronic Theses and Dissertations at Robert Gordon University in Aberdeen, Scotland, on June 4, 2008. This half-day preconference workshop will

---

19 Katherine Skinner (kskinne@emory.edu)
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 services to and responsibilities of each institutional membership category.

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 Joint Conference on Digital Libraries (JCDL) on June 20, 2008, in Pittsburgh, Pennsylvania, at the Omni William Penn Hotel. Typically the all-day MetaArchive workshops provide information and training for institutions seeking to build or join LOCKSS-based distributed digital preservation networks.

14. Reports

Members of the MetaArchive Cooperative have access to many reports that are generated programmatically and available online through the Conspectus Database. Individualized reports are distributed annually to ETD Archive members and document a number of milestones:

15. Replacement available in the ETD Preservation Network

If NDLTD institutions are Preservation or Sustaining Members, their caches will have automatically replaced ‘bad’ files from the ETD Archive. When Contributing Members find it necessary to replace lost or corrupted ETDs in their local collections, they can request replacements from the ETD Archive by contacting the MetaArchive Program Manager.

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

---

20 http://metaarchive.org/JCDL_Workshop
21 Katherine Skinner (kskinne@emory.edu)