ETD PRESERVATION SURVEY RESULTS
MetaArchive and NDLTD Collaborate to Provide a Distributed Preservation Network for ETDs

Gail McMillan
Digital Library and Archives, Virginia Tech

Abstract

Because many universities now welcome or require ETDs from their graduate students, institutions must ensure that these works will be at least as available and enduring as they were when libraries and archives preserved the bound print volumes on their shelves. To this end, the Networked Digital Library of Theses and Dissertations sponsored an online survey to help gauge the digital library community's interest in a distributed digital preservation network (DDPN) specifically for ETDs. Over 90 institutions responded to the survey in early 2008, including more than one-third who heard about it from the NDLTD and ETD listservs. Based on the enthusiasm expressed in the survey, the MetaArchive Cooperative (www.metaarchive.org), which successfully deploys a DDPN among six diverse universities in the southeastern United States, is opening the Cooperative's services and resources to the NDLTD. This paper describes survey responses and aspects of the NDLTD Preservation Strategy.

Introduction

Essentially all theses and dissertations created today are born-digital and 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 research and scholarship that comprises this important component of our institutional histories.

Digital preservation is the systematic management of computerized information over an indefinite period of time. It demands continual attention and this constant input of effort, time, and money to handle changes in technology and organizations is the main obstacle to preserving digital information beyond a few years. Effective preservation succeeds by replicating copies of digital content in secure, distributed locations over time because security reduces the likelihood that any single cache will be compromised and distribution reduces the likelihood that the loss of any single cache will lead to a loss of the preserved content. A single organization is unlikely to have the capability to operate several geographically dispersed and securely maintained servers. Inter-institutional agreements must be put in place or there will be no commitment to act in concert over time.[1]

The Networked Digital Library of Theses and Dissertations (NDLTD)[2] and the MetaArchive Cooperative[3] share the goal of helping higher education institutions provide long-term open access to ETDs. The MetaArchive Cooperative is a service organization whose mission is to support, promote, and extend the practice of distributed digital preservation. The MetaArchive and the NDLTD joined forces in 2008 to offer preservation services for ETD collections by implementing an ETD Archive using the technological approach called distributed digital preservation network (DDPN). Participants in this new archive make their collections available for harvesting into the network and they may also participate in the Cooperative by hosting a LOCKSS-based networked server

[1] Halbert, Martin. “MetaArchive” presentation to SCHEV LAC, March 28, 2008.
[2] http://www.ndltd.org/
[3] http://www.metaarchive.org/

LOCKSS (Lots of Copies Keep Stuff Safe) is an international non-profit community initiative that provides tools and support so libraries can easily and cost-effectively preserve today's web-published materials for tomorrow's readers.[4] Typically the LOCKSS open-source software programmatically collects content from publishers and distributes copies among partner libraries' inexpensive servers where it is preserved The software also audits and repairs content as needed from the publisher or the partners. It allows content to be disseminated only to the appropriate users and the host library's clientele see the content from the publisher's site, unless it is not available from there. Then it is served from the partners' copies, which otherwise are used only to audit and repair the digital content. Many libraries are familiar with this simple, robust, low maintenance, low cost distributed digital preservation system.

In 2004 six American university libraries received funding from the Library of Congress to create a similar network of trusted partners and adapt LOCKSS by disconnecting access from preservation so that the partners' servers become a networked secure dark archive. This partnership became the MetaArchive Cooperative and four years later it is ready to expand its network to include those members from the NDLTD who also seek a tested and effective preservation strategy familiar to libraries. Collections of born-digital and digitized theses and dissertations from NDLTD institutions will be ingested into the ETD Archive by the MetaArchive system and copied, distributed, and stored on secure servers at multiple NDLTD partner institutions. The MetaArchive Cooperative will not provide access; that service will remain with each institutional member hosting an ETD collection.

To determine if there was, indeed, a desire for an ETD-specific preservation network, the MetaArchive Cooperative, in consultation with the NDLTD, designed an online survey with 14 multiple-choice and short answer questions that Virginia Tech's Digital Library and Archives' hosted.[5] The NDLTD Board of Directors received preliminary survey results at its January 21, 2008 meeting, and voted to endorse a distributed preservation network for ETDs within the MetaArchive. This paper describes the survey responses and aspects of the NDLTD MetaArchive preservation strategy for the ETD Archive.

Various academic listservs announced the ETD preservation survey, resulting in 95 completed surveys as of April 10. Below is the summary of the sources of those responses to the call for participation.

How did you learn about this survey?

	Listserv Sources of Survey Responses
9%	Council of Graduate Schools
10%	Digital Library Federation
11%	Association of Southeastern Research Libraries
15%	Association of Research Libraries
23%	Other
32%	NDLTD and ETD

Because of the significant portion of the survey respondents' who are members of the NDLTD and ETD listservs, and because this discussion has been prepared for the NDLTD-sponsored conference, this essay groups their responses together and refers to them as ETDL. When noteworthy this paper highlights and/or contrasts them to the non-ETDL, that is the responses from the other listservs.[6]

[4] http://www.lockss.org/
[5] http://lumiere.lib.vt.edu/surveys/
[6] This level of analysis was enabled by the skills Kimberli Weeks, online editor at Virginia Tech's Digital Library and Archives.

At least half of the ETDL are current members of the NDLTD. Of these, nearly one-third of the respondents were at international universities, a little more than half were at American universities, and less than one-fourth of the respondents were at undesignated institutions.

Survey Responses

Over three-fourths of the survey responses came from universities that accept ETDs. Over one-third of those respondents also reported that their institutions accept just the electronic formats while just over half of them also maintain print copies. It was expected that institutions with active ETD initiatives would be keeping up with relevant issues through the NDLTD or ETD listservs so it is not surprising that more of those universities accept ETDs. However, a smaller percentage of those institutions accept only electronic versions.

All Surveys		ETDL	Non-ETDL
80%	Accept ETDs	84%	77%
39%	Accept Electronic Only	16%	24%
57%	Maintain Print Copies	22%	35%

File Formats of ETDs

Each institution in the MetaArchive Cooperative provides detailed descriptions of their preservation collections, and this metadata is stored in the MetaArchive Conspectus Database.[7] The metadata is currently used for a variety of purposes including network administration but it also has anticipated future uses such as format migration. The MetaArchive, like LOCKSS, is format agnostic, ingesting all file formats into the DDPN. However, the survey asked respondents to select from among 16 text, image, audio, and video file formats which ones were accepted with ETDs[8] because this is an important element of preservation planning and format migration considerations.

Taking into consideration that the any-format option may dilute the responses for any specific file format, it is still worth noting the formats selected most frequently as noted in the chart below.

What file formats do you support for your ETDs?

File Formats in ETDs	
85%	PDF
30%	JPG
27%	WAV
24%	GIF
23%	HTML
23%	MOV
21%	AVI
21%	MP3

[7] http://www.metaarchive.org/conspectus/
[8] See the NDLTD list of recommended file formats at http://etd.vt.edu/howto/accept.html

More ETDL accept QuickTime movies (4/5), XML files (7/12), and PowerPoint slides (4/7), while non-ETDL more often selected the audio formats WAV (12/19) and MP3 (9/15), AVI videos (9/15), and JPG images (13/21). Other formats listed by survey respondents included MIDI, CVS, TXT, and JP2. The range of file formats that comprise ETDs at the survey respondents' universities, matches those the MetaArchive has already ingested and those that were damaged and repaired during extensive DDPN tests.

Platforms and Repositories

The current MetaArchive members have experience with a variety of platforms and repositories, and they have begun to prepare recommended best practices for organizing ETD collections to facilitate harvesting and ingesting into the preservation network. Virginia Tech has prepared the guidelines for collections created with ETD_db[9] and other members are documenting their work with CONTENTdm (Auburn), DSpace (Georgia Tech), and Fedora (Emory).

The survey anticipated that institutions were using a variety of platforms and popular repository structures to collect, disseminate, and/or store ETDs. As the table below reports, the survey responses confirmed that the majority of ETDs are not uniformly part of any particular platform or repository, and that a significant portion are in home-grown systems and in unanticipated repositories. Among the others listed were CONTENTdm, DigitalCommons, DigiTool, and ProQuest.

What platform or repository structure are you using to collect, disseminate, and store your ETDs?

All Surveys	Platforms/Repositories with ETDs
29%	In House system
29%	Other
26%	DSpace
13%	ETD_db
3%	Fedora
1%	EPrints

Collection Structures

Asked how institutions currently organize their ETD collections, seven categories came to light, although 67% use one of three structures: subject, year, or everything-in-one-collection. Of these categories, most frequently ETDs were organized by subject-like categories according to departments, colleges, or disciplines, according to 25% of the responses. Tied for the next most mentioned collection organization with 21% each, were everything-in-one-collection and collections based on the year the degree was granted. Three other categories mentioned were accessibility, degree, and author

Open Access and Dark Archives

When the NDLTD Board of Directors was considering the collaboration with MetaArchive for the ETD Archive, several members stressed the importance of open and unimpeded access to ETDs. The MetaArchive also believes that ETDs should be openly accessible, but that access should come directly from the authors' home institutions rather than from the ETD Archive.

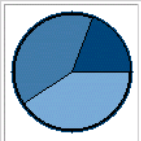
[9] See "NDLTD MetaArchive Preservation Strategy" http://scholar.lib.vt.edu/theses/preservation/

The MetaArchive's DDPN diverges significantly from the LOCKSS principle of open access from the preservation network. The current NDLTD MetaArchive preservation strategy separates the preservation archive from public web access, resulting in the practice of dark archiving where all ETDs are completely inaccessible to any server outside the specifically designated preservation partners in the network.

While considering future development of the MetaArchive Cooperative, its Steering Committee acknowledged that some changes would be necessary in order to evolve to meet the needs of potential members. Future development of the MetaArchive Cooperative may include adopting the LOCKSS feature that enables public access from the preservation network but it would be each member 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 became inoperable.

If it is to attract new members from among the majority of universities that responded to the ETD preservation survey, the MetaArchive may indeed have to reconsider its stand on the separation of access from preservation. More than three-fourths of the survey respondents preferred accessible preservation archives as the table below illustrates.

If [your institution would be interested in participating in an ETD-specific LOCKSS-based collaborative distributed digital archive sponsored by the NDLTD] would there be a preference for



18.99%	Dark archiving (i.e., inaccessible to the public)	(15/79 responses)
40.51%	Open archiving with public access	(32/79 responses)
40.51%	Dim archiving (i.e., designated files inaccessible but others available to the public)	(32/79 responses)

LOCKSS Distributed Digital Preservation Network

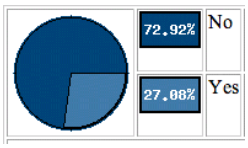
More than two-thirds of the survey respondents reported having experience with or knowledge of LOCKSS-based preservation networks. All but eight percent of the survey respondents said that their institutions would be or might be interested in participating in a LOCKSS-based ETD archive sponsored by the NDLTD. Many more non-ETDL said "maybe."

Would your institution would be interested in participating in an ETD-specific LOCKSS-based collaborative distributed digital archive sponsored by the NDLTD?

All Surveys		ETDL	Non-ETDL
42%	Yes	20%	22%
49%	Maybe	15%	35%
8%	No	5%	3%

Lack of Preservation Planning

The most surprising response to any question in the ETD preservation survey was the response to "Does your institution have a formalized preservation plan for its ETDs?" Only about one-quarter of the universities responding indicated that they have a preservation plan for their ETDs, leaving nearly three-fourths of the universities that accept ETDs without formal preservation plans.



Does your institution have a formalized preservation plan for its ETDs?

Correlating responses for universities that accept ETDs with those that have formal preservation plans reveals that less that one-fifth (18%) of the ETDL with ETD collections also have formal plans. Two-thirds of the ETDL accept ETDs without having formalized preservation plans. Only one institution has a formal plan but does not accept ETDs.

With such a small percentage of universities indicating that they have preservation plans for their ETD collections, this survey reached a significant number of institutions that could clearly benefit from joining the NDLTD MetaArchive preservation network.

Participating in the MetaArchive Cooperative

The survey sought to determine if there was interest in not only preserving ETD collections, but also in having a role in the preservation activities as participating members of the MetaArchive Cooperative. There are three membership categories in the Cooperative, with institutional participation ranging from minimal to considerable. MetaArchive Cooperative Charter[10] fully describes the benefits as well as the obligations of each membership category.

If [your institution would be interested in participating in an ETD-specific LOCKSS-based collaborative distributed digital archive sponsored by the NDLTD] what level of participation might your institution support?

All Surveys		ETDL	Non-ETDL
46%	Contributing Membership	42%	49%
30%	Preservation Membership	29%	30%
24%	Sustaining Membership	29%	21%

Contributing Members

Institutions that join the MetaArchive Cooperative as Contributing Members contract for services only and do not have responsibilities beyond preparing their own ETD collections for harvesting. These institutional members do not have any technical obligations, nor do they have an active role in the operation of the Cooperative. The MetaArchive prescribes how Contributing Members organize their ETD collections to facilitate harvesting and ingest into the ETD Archive. Contributing Members are allocated five gigabytes for their ETD collections, though they may [insert *opt*] to purchase additional space.

Nearly half of the survey respondents indicated that their universities would be interested in being Contributing Members, that is, they would contribute to the ETD preservation network by making their collections available for harvesting but they would not operate a node on the network. Nearly half of the non-ETDL favored of this level of participation.

[10] http://www.metaarchive.org/pdfs/MetaArchiveCharter_0707.pdf

Preservation Members

Institutions willing and able to share DDPN responsibilities could join the Preservation Members category. These universities will operate a node on the ETD preservation network, preserving not only their own ETD collections, but also ingesting those of at least six other NDLTD institutions. Preservation Members are required to maintain servers and network nodes that meet the MetaArchive's specific technical requirements.[11] The preservation nodes collectively comprise the distributed MetaArchive preservation network.

Nearly one-third of the survey respondents indicated that their institutions would like to actively participate in the MetaArchive Cooperative as Preservation Members, archiving their ETDs in the distributed network, running a secure server for the network, and harvesting and caching ETDs for other NDLTD members. Preservation Membership comes with 20 GB of storage for their collections, though institutions may purchase additional space.

Sustaining Members

The MetaArchive Cooperative's Sustaining Members have the most responsibilities and the greatest opportunity to affect the Cooperative (fully described in the Cooperative's Charter[9]). Along with the responsibilities of Preservation Members, Sustaining Members also develop and test software, networking, and transmission standards, and they research and deploy the work of the Cooperative, contributing staff and resources. Sustaining Members each receive 40 GB of archiving space in the DDPN though they may purchase additional space.

Nearly one-fourth of the survey respondents indicated they wanted to join the MetaArchive as Sustaining Members. While nearly one-third of the ETDL selected this category, it was of less interest to the non-ETDL.

Open-ended Comments

Of the 95 completed ETD preservation surveys, 65 institutional representatives responded to open-ended questions. They provided useful information for the MetaArchive Cooperative to consider about the concerns of potential members of an ETD preservation network.

Information Needed before Joining the MetaArchive

One of the final survey questions offered the opportunity to suggest what would help institutions make informed decisions about whether to participate in the ETD preservation network. Thirty-five categories grew out of the 65 narrative responses. The largest number of comments had to do with financial concerns—slightly more than half wanted to know about the costs involved in the MetaArchive Cooperative's DDPN. The second most-mentioned information need came from nearly one-fourth of the respondents who wanted to know more about human resources, including skills; followed by hardware and platform concerns mentioned by nearly one-fifth; then concerns about responsibilities, expectations, requirements and technical issues; and finally ETD access, policies, and procedural concerns. [The survey continues to be available online and readers of this paper may consult the most current data.[12]]

[11] http://www.metaarchive.org/pdfs/AppendixA0208.pdf
[12] http://lumiere.lib.vt.edu/surveys/ Select Digital Preservation of ETDs from the drag-down list.

The largest number of comments, made by half of the ETDL, from most to least, was cost, human resources, and access to ETDs. Nearly one-third mentioned costs; far behind but echoed by 13% were human resources, and access to ETDs was mentioned by 7%. The remainder was a varied list of 23 comments, including hardware/platform, national networks, requirements, work to prepare archives, documentation, liability, policies and procedures, security, support, and sustainability. Comments that were made by the ETDL, but were not mentioned by the non-ETDL, included documentation and national networks.

The largest number of comments, made by nearly half of the non-ETDL, from most to least, was cost, hardware, technical issues, and human resources. Their most frequently mentioned concern was also about costs, made by 20%. Half as many were concerned about policies and procedures, followed by hardware, platform and other technical issues. Fewer wanted to know about human resources/staff/skills and general responsibilities and requirements.

The dichotomy of responses is sometimes quite striking. Non-ETDL mentioned technical issues more than twice as often, while ETDL more often mentioned human resources.

Comments and Concerns about an ETD Preservation Network

The final survey question asked for comments and concerns about preservation of ETDs, particularly the distributed model that the MetaArchive Cooperative offers. The most common comment, made by nearly one-fifth of the respondents, was that this preservation strategy provided a welcome opportunity.

- A welcome opportunity for academia to regain control of its intellectual properties, and cost saving through cooperation.
- We believe that it is a very useful solution, especially for institutions that do not have specific and formalized preservation plan[s].

The second most-made comments were about functionality, including format migration. Examples include

- I'm very concerned about how to migrate materials to archival quality formats while enabling students to work with a variety of formats.
- I would like to know more about your logic preservation strategies (renderable, readable) for multiple types of formats.

Comments on the current limitations of LOCKSS got the second-to-the-largest percentage of comments from the ETDL. These included confidentiality, format migration, and improving functionality. The non-ETDL did not mention concerns about confidentiality. Their top two concerns were about their not knowing enough and about the robustness of the DDPN strategy. These institutions also commented on issues that the ETDL did not, including repository software.

Conclusion

In January 2008 Virginia Tech's Digital Library and Archives posted an online ETD Preservation Survey designed by the MetaArchive Cooperative and sponsored by the NDLTD. By April 10, 95 institutional representatives had completed the survey, which they had learned about through academic listservs aimed at library and graduate school leaders as well as other members of the higher education community. The

survey responses indicate that any preservation strategy must accommodate a range of standard file formats, a variety of repository systems, and ETD collections using various organizational structures.

The surveys indicate that ETD collections exist at 80% of the universities responding. However, 74% of those universities lack formal preservation plans for their ETD collections. This chasm demonstrates the dire need for a preservation strategy such as that offered by the MetaArchive Cooperative. A distributed digital preservation network may be a particularly good fit for an international organization such as the NDLTD as well as the fact that two-thirds of the responding universities already have experience with LOCKSS. An ETD-specific LOCKSS-based collaborative DDPN sponsored by the NDLTD is of interest to the majority of survey respondents.

Not only did this survey demonstrate the need for and interest in a formal preservation strategy for ETDs, the majority of survey respondents also want to participate in the preservation activities, not just off-load their ETDs into a secure archive maintained by others. If this interest in preservation network participation becomes linked with the respondents' interests in a more open ETD Archive, it will cause the MetaArchive to reexamine one of its founding principles—the separation of access from preservation in the distributed archive. The enthusiastic response combined with the number who might be interested in joining the MetaArchive Cooperative, especially at the more participatory levels, was a welcome outcome of the survey for the MetaArchive Steering Committee. Finally the NDLTD can provide guidelines for long-term access and preservation of ETDs.