CFP: Call for preservation!

Results and discussion of a survey on the preservation of electronic theses and dissertations

Kelsey Libner
North Carolina State University Libraries
kelsey_libner@ncsu.edu
NCSU Libraries, Campus Box 7111, Raleigh, NC 27695-7111 / USA
www.lib.ncsu.edu/dli/
Keywords: Preservation, preservation policies, digital preservation, format, migration, survey, backup, PDF

Abstract

16 institutions were surveyed regarding practices for long-term preservation of ETDs. The first part of this paper reports results of the survey. Results are organized into topical areas including: electronic file format; electronic file backup; physical form of backup, if any; differences in the processing of master's theses vs. dissertations; and migration. Policies and practices vary widely across institutions and are in many cases a work in progress. The second part of the paper takes up questions surrounding file format of both the main document and supplementary files. The structured presentation and discussion of survey results is indended to underscore the complexity of the preservation challenges that lie ahead and to provide guidance in framing future discussions among librarians, archivists, and faculty.

Acknowledgments

I would like to acknowledge and thank North Carolina State University for its support of both the survey on which this paper is based and my travel to ETD 2003.

Introduction

> "We really do need to do something on the preservation side. It's a little embarrassing to admit but we'll get there."
> Preservation policies are "definitely a work in progress and we are closely following the emerging standards..."
> "Lots of thoughts and nothing definite."
> - Three survey respondents commenting on their institutions' ETD preservation policies

Theses and dissertations represent a significant investment of time by graduate students and their supervising faculty. Collectively, they may be viewed as an important part of the intellectual heritage of an institution. These are compelling reasons to ensure their long-term preservation. But questions arise as to how best to preserve electronic theses and dissertations (ETDs) which are "born digital." For many printed documents a policy of "benign neglect" suffices for preservation (Teper and Kraemer, 2002): put them on the shelf and make sure there's a roof above. Barring floods, fires and tropical humidity, most of your materials will be preserved.

By contrast, preservation of digital documents is an active process. Hardware, operating SYSTEMs, software and text encoding standards are evolving rapidly. If no measures are taken to ensure access to a document in each new generation of computing environments, it will effectively be lost. In the words of Besser and Lyman (1998): "Without intervention, the default condition of paper is persistence; the default condition of electronic signals is interruption."

"An active program of digital preservation will require a significant investment over the long term in planning, implementation and long-term supervision; however, Beagrie and Jones (2002) warn: ""The costs of recreating a digital resource may be much higher than those for preserving it; further, the opportunity to do so may no longer exist." Serious engagement with questions about the digital preservation of ETDs is one way in which research libraries can continue to develop, in a digital context, their role as trusted stewards of intellectual work.

This paper does not provide a step-by-step guide to creating a preservation program for ETDs. Instead, it is hoped that a structured discussion of a survey on ETD preservation will underscore the complexity of the preservation challenges that lie ahead, and prove helpful in framing future discussions among librarians, archivists, and faculty.

A survey of ETD preservation at 16 institutions

Purpose and method

In September and October of 2002 a survey on the preservation of ETDs was carried out.[1] A list of institutions with either the option or the requirement to submit an ETD *in fulfillment of graduation requirements* was gathered. Policies of the following institutions were examined:

- American universities in the Digital Library Federation

- North American universities listed as official members of the Networked Digital Library of Theses and Dissertations (NDLTD).
- Association of Research Libraries (ARL) member institutions ranked above NCSU Libraries (i.e., 31 and above in the 2001 ranking).

We learned of additional institutions through an inquiry on the ETD-L listserv. Our final list included 17 institutions (NCSU is not included in this list or in the results below).

Through Web research and phone calls, individuals familiar with ETD preservation practices and workflow were identified at each institution and asked to participate. We offered to share aggregate results of the survey with participants. The survey was administered by phone or e-mail, depending on each participant's preference, in September 2002.

We received responses from every targeted institution except one. The 16 survey participants are listed in Appendix A. The questions asked in the survey are attached as Appendix B.

Results

Results are presented in the following topical areas: general ETD policy, electronic file format, electronic file backup, paper copies, physical copies of dissertations, differences in processing master's theses and dissertations, migration plans, and workflow and coordination. Because of space limitations, some results are summarized without further discussion.

As agreed with survey participants, responses are reported without attribution to particular individuals or institutions.

General ETD policy

Dissertations in electronic form: Required at 11 institutions; optional at 5 institutions.

Master's theses in electronic form: Required at 7 institutions; optional at 6 institutions; not an option at 2 institutions.[2]

Electronic file format

- 9 of 16 institutions require PDF for the main document.
- 7 of 16 accept PDF or other formats (mainly HTML) for the main document.
  - 3 institutions accepting HTML either have strict guidelines for it or convert to PDF.
  - Nonstandard formats accepted include Postscript (1 institution) and XML or XHTML (1 institution)

File format of the main document

Seven respondents expressed reservations about PDF. At the same time some respondents recognized the value of PDF and the barriers to implementing an XML-based ETD program. Their responses follow:

> "[Some on our ETD committee] want to move to something that's more flexible, something that can be migrated more easily than with PDF... Acrobat 5 is supposed to be a little more flexible where you can save as HTML or XML. I'm not terribly excited about it."

> "We're hoping [the use of PDF] is a temporary situation... I don't want to be beholden to Adobe."

> "As soon as export to XML export works we'll try to do that... We assume that PDF will be going away and that we will be converting to XML sometime but not at the present time. We're committed as much as possible to open source standards."

> "...There don't seem to be too many easy answers. PDF is easy for the student to produce and to give to the library but it's difficult to believe that the library could sustain that for any amount of time. On the other hand creating HTML or XHTML puts an enormous burden on the student and they would likely need support from the library that we at least now don't feel that we could provide. Although the end result would be more sustainable it requires a whole magnitude of effort from the student and therefore from the library as well to support it. So being able to influence which format is best is a difficult question to answer."

> "We're looking forward to SGML or XML as being a format for both preservation and access but we're not going to ask our students to put it into XML until there's an editor [i.e., software to simplify the process of saving in XML format]."

> "...In terms of people doing full-blown XML, the time has not been ripe, particularly at the student level. [The current SYSTEM is] already like pulling teeth with some students. There's a learning curve on producing XML documents. [XML features are] just coming out on WordPerfect and MS Word. But it's not an easy thing. We know that this holds a lot of promise for the future - there are all the goodies that go along with XML including migration. If XML becomes the technology of the future, we should be able migrate our

1 The survey was conducted to provide background for a review by North Carolina State University (NCSU) Libraries of its own practices in managing and preserving ETDs. This review was in part prompted by the requirement, established in July 2002, that all NCSU theses and dissertations be submitted in electronic form.

2 One institution is not included in this count because in general it does not offer or require master's theses.

collection. This is a leap of faith, but we took a leap in 1998 [by starting to accept ETDs]."

"We realize that this format [PDF] is not going to last forever."

SUPPLEMENTARY FILES

Four respondents expressed concerns about supplementary files. One said that if authors are "attaching multimedia and getting fancy with their supplementary files they ought to do it at their own risk." Another said that "The document should work without ancillary materials... [These are] just an enhancement." If supplementary files fail, it was explained, the main document can still stand alone as a complete and coherent thesis or dissertation. The ETD website of a third institution states: "The advantages of access must be weighed against the value of long-term preservation of textual content. The candidate and supervisor may wish to consider limiting multimedia to appendices that *enhance but are not required* for comprehension of the thesis" (emphasis added). The fourth institution simply does not accept files in any format other than PDF: "We don't like the multi-files, trying to sort them out. We're trying to go with a very high-quality PDF that is appropriately backed up and preserved."

Electronic file backup

Major backup measures reported by survey participants:
1. General strategies:
- Inclusion of ETDs in existing backup plan for digital publications and/or library files
- Use of statewide or national entity for backup (e.g., Florida Center for Library Automation, OhioLINK, National Library of Canada, WVNet)
- Electronic storage at UMI[3] as a component of the institution's backup plan

2. Specific measures taken:

- Regular copying from production server to backup server
- Backup to two off-site locations
- RAID ("Redundant Arrays of Independent Disks") storage
- Weekly or nightly backup to tape (with regular replacement of tapes)
- Off-site storage of tapes
- Backup to CD annually or each semester
- Storage of CDs in a safe or fireproof space.

Paper copies

2 institutions are using a paper copy for archival purposes.

4 institutions mention the creation of paper copies for non-archival purposes (as part of the review process or as required by departmental libraries, departments, or committees).

10 institutions make no mention of paper copies.

Physical copies of dissertations

7 institutions require submission to UMI (which generates and retains a microfiche copy of the file) but don't keep a local physical copy.

7 institutions, in addition to requiring submission to UMI, keep a local physical copy or have access to a non-UMI physical copy (5 microfiche, 2 paper).

2 institutions do not require submission to UMI and are not retaining a paper or microfilm copy.

One respondent remarked, "I really view output to microfilm as an enormous plus in terms of digital preservation. That's a luxury we do not have for most of our other digital publications."

Processing of master's theses vs. Dissertations

Where both theses and dissertations are processed, 7 institutions process each in the same way for preservation purposes, while 5 have weaker preservation measures in place for theses.

Migration plans

7 institutions reported stronger positions or actions.
6 institutions reported general intentions.
3 institutions had no policies or plans to report.

Migration-related plans include:
1. Making a commitment to future migration of a specific file type or types as necessary
2. charging students $10 per document (over and above UMI processing fee) for future costs to maintain access
3. making an inventory of file formats "as a first step"
4. delegating migration responsibility to a government-sponsored library agency (e.g., National Library of Canada; interest expressed regarding Florida Center for Library Automation).

Workflow and coordination

Three institutions pointed to the importance of workflow and coordination in managing their ETD program:

"The effort of bringing collections online is integrated into the entire library effort. All departments are involved. Certainly the provost's office and the administration have charged the libraries to preserve the [ETDs]."

There has been a lot of "high-flying rhetoric around ETDs but ours is a very simple SYSTEM. Others have used our model. Students and faculty understand it. We have a good workflow for

3 URL: *http://www.umi.com/*

doing format checks. People were freaking out. Once they figured it out, it has run smooth as silk."

"The provost has been really instrumental. The archival and administrative issues are important. Support from the top has been vital."

Discussion of selected results

File formats: A closer look at Portable Document Format (PDF)

PDF and Adobe Systems Incorporated

Adobe's commitment to maintaining PDF as an open and published standard mitigates concerns about its copyright on the standard. However the format should not be viewed, necessarily, as the format of record over the very long term. Mark Ockerbloom writes:

> As data formats go, PDF is particularly likely to be supported for a long time, and to spawn migration paths... Even so, it is likely that PDF will one day be superseded by another format. It may be a successor format (as PDF is to Postscript), or it may be a completely different format that users prefer over PDF. Hence, it is necessary to have migration strategies planned for PDF. (Ockerbloom, 2001)

Despite these long-term concerns, PDF is valuable in that it provides a relatively faithful rendering of a page and document across a range of platforms. To borrow the phrase of Michael J. Patrick of Ansyr Technology Corporation, PDF has good "paper fidelity". At the same time, because the imaging model describes contents in an abstract way[4], the format allows much more than simple rendering of page images. It allows text search; integration of multimedia objects; hyperlinks within and outside the document; access to content using alternative reading devices; and, using Adobe InDesign, export to XML. PDF may also contain raster images such as TIFF files. With the ability either to render pages from abstract descriptions or display flat page images, PDF works - for the moment - as a convenient transitional format from theses and dissertations on paper to those in digital form.

In search of an archival PDF format

The underlying model of PDF, based on the PostScript Page Description Language, has been relatively long in development and is not likely to change significantly. This foundation of the format seems fairly solid from a preservation perspective. It also offers the possibility of future export to another format. However recent changes to the specification concerning annotation, highlighting, digital signatures, and object transparency - plus the possibility of future changes - raise concerns about using PDF as the format of record.

A committee under the joint auspices of The Association for Suppliers of Printing, Publishing and Converting Technologies, and the Association for Information and Image Management, International has been formed to develop a standard archival format of PDF called PDF/A. The stated goal is to "develop an International standard that defines the use of the Portable Document Format (PDF) for archiving and preserving documents" (NPES/AIIM, 2002). Unless and until a viable archival-PDF standard emerges, preservation plans must account for the limitations of PDF as we know it - what we might conservatively call *non-archival* PDF.

Archival storage of the main document

In a paper based on a presentation given at the ARMA 2000 conference, Steve Gilheany argues for the retention of a range of different formats: the native-format document, the vector-based PDF document, and TIFF files (Gilheany, 2000, cited in Teper and Kraemer, 2002). Each has advantages and disadvantages:

Table 1: Advantages and disadvantages

Format	*Pros and cons*
Word, LaTeX or other native format	Pro: Can be edited. Because it is the original form of the document, it does not contain conversion anomalies. Con: Relatively short lifespan.
PDF-vector	Pro: More durable than Word format. Contains machine-readable text allowing search, document rendering on multiple devices, and greater accessibility than TIFF files. Con: Subject to changes in PDF specification. Migration may change document formatting.
TIFF files	Pro: Extremely simple and durable. A de facto standard for digital masters (see Kenney, Rieger, & Entlich, 2003). Con: File is "flat", sacrificing functionality such as non-image multimedia files. Larger file size.

4 "A high-level imaging model enables applications to describe the appearance of pages containing text, graphical shapes, and sampled images in terms of abstract graphical elements rather than directly in terms of device pixels." (Adobe Systems Incorporated, 2001, p. 10).

Questions to consider

- Are students using non-established features of PDF such as annotations and digital signatures? If so, what is the preservation impact of these decisions?
- If the PDF/A standard continues to develop, would it be a candidate for use by ETD programs?
- Is the cost of maintaining multiple, redundant formats (including TIFF page images) justified based on the priority of preserving ETDs?

File formats: Policies on supplementary files

Anecdotal evidence from the survey suggests that the percentages of ETDs submitted in formats other than PDF, HTML, and JPEG run quite low (for one institution, about two percent). Policies on supplementary files can be put into two general classes, "conservative" and "liberal". Under a "conservative" policy, a limited number of alternative file formats is accepted but a strong commitment is made to preserving all files. Under a "liberal" policy a broader range of file formats is accepted but the preservation commitment may vary for different formats.

Questions to consider

- Will supplementary files come to play more of a substantive role in theses and dissertations? If so, what are the implications for preservation policy and practice?
- Should an institution pursue a conservative or liberal policy on file formats? One option to consider is to specify differing levels of commitment to preserve and migrate depending on format (see, for example, the model used by the Harvard Digital Repository Service[5], specifying different levels of commitment for different file formats.)
- Is the ETD genre evolving from a static "presentation model" towards a more flexible genre with room for alternate formats - a hybrid of page presentation and functionality? The graduate school administration and the library (or other preserving entity) may wish to take up this question together since it has implications both for graduate school policy and long-term preservation.

Conclusion

These results could easily extend to discussion of: the importance of making an explicit statement of preservation commitment; the usefulness of a lifecycle view in digital preservation planning; the challenge of modifying or creating new organizational routines in response to digital preservation challenges; strategies for controlling the costs of digital preservation; the use of preservation metadata schemes such as METS for ETDs; and ETD preservation in the context of general models and infrastructures for digital preservation, including the OAIS Reference Model and institutional repositories. While I do not have space to discuss these topics, I hope the results and discussion above provide some basis for considering the connections.

Digital preservation presents a host of new problems and contexts; however, at a fundamental level it remains a familiar challenge: for libraries and archives to act as responsible and effective stewards of intellectual heritage. Preservation programs are most likely to succeed when the mundane organizational and technical concerns have strong connections to the fundamental role of stewardship. The above results and discussion are offered with an interest in suggesting and strengthening these connections.

5 See the Policy Guide of the Digital Repository Service at *http://hul.harvard.edu/ois/SYSTEMs/drs/policyguide.html*.

Survey Participants

Table 2: Survey Participants

Institution	Participant	Title
California Polytechnic Institute	Eric F. Van de Velde, Ph.D.	Director of Library Information Technology
California Polytechnic Institute	Kimberly Douglas	Director of Engineering and Applied Science library
East Tennessee State University	Wesley Brown, Ph.D.	Dean, School of Graduate Studies
The Ohio State University	Tim Watson	Director, Graduation Services, Graduate School
Pennsylvania State University	Sue Kellerman	Judith O. Sieg Chair for Preservation (Head)
University of Florida	Martha Hruska	Director for Technical Services
University of Georgia	Susan Gants	Management information specialist, Systems Department, University of Georgia Libraries
University of Georgia	David K. Knox	Director of Information Technology, The Graduate School
University of Kentucky	Beth Kraemer	Electronic Resources Librarian
University of Pittsburgh	Tim Deliyannides	Head of Department of Information Systems, University Library System
University of South Florida	Monica Metz-Wiseman	Coordinator of Electronic Collections, Collection Management
University of Tennessee	JoAnne Deeken	Head of Technical Services and Digital Access
University of Texas at Austin	Dennis Dillon	Assistant Director for Collections and Information Resources
University of Virginia	Melinda Baumann	Director of Digital Library Production Services
University of Waterloo	Christine Jewell	Head, ILL and Document Delivery
Virginia Polytechnic Institute and State University	Gail McMillan	Director, Digital Library and Archives, University Library
West Virginia University	John Hagen	Library Associate and ETD Coordinator, Acquisitions Department
Worcester Polytechnic Institute	Helen M. Shuster	Director of Library Services (within IT division)

Survey questions

The survey was given over the phone or via e-mail (depending on preference) to the individuals listed in Appendix A.

Part I. Policy on paper and electronic dissertations

1. Please confirm: At your institution, are ETDs accepted as the official version of a thesis or dissertation? If so, is the ETD optional or required?
2. Is a student submitting an ETD also required to submit a paper copy to:
   A. the graduate school? (yes or no)
   B. the university library? (yes or no)
3. Do the policies for master's theses and dissertations differ? If so, how?
4. What electronic formats are acceptable for the thesis or dissertation? (e.g., PDF, HTML) (A summary would be fine; an exhaustive list is not necessary.)

Part II. Preservation policy or arrangements

5. At your university, what entity is responsible for archiving theses and dissertations?
6. Please describe the policy or plan for preserving ETDs. We're particularly interested in the following:
   A. Do you also keep a paper and/or microfilm copy of ETDs? If so who creates these copies?
   B. Any plans, policies, or thoughts on:
      i.migration of file formats
      ii.backup plan, including offsite storage of ETDs, and ensuring data and file integrity
7. Briefly, what is the policy for access to ETDs? (Where made available?) Are students able to restrict access? If so, what kind of restrictions are permissible?
8. Would you be willing to share any documents on your policy or plan? If so, please attach them or provide a URL here.
9. Do you have any other thoughts on the preservation of ETDs at your institution or in general?

References

Adobe Systems Incorporated (2001). *PDF Reference, Third Edition: Adobe Portable Document Format Version 1.4*. Boston: Addison-Wesley. Retrieved February 19, 2003 from *http://partners.adobe.com/asn/developer/acrosdk/docs/filefmtspecs/PDFReference.pdf*

Beagrie, Neal and Maggie Jones (2002). *The Preservation Management of Digital Material Handbook*. Retrieved March 3, 2003 from *http://www.dpconline.org/graphics/handbook/*

CCSDS Secretariat (2002, January). *Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, BLUE BOOK*. Retrieved on February 24, 2003 from *http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf*

Gilheany, Steve (2000). Permanent Digital Records and the PDF Format. A white paper based on a presentation at the ARMA 2000 conference (Association of Records Managers and Administrators, International). Retrieved February 24, 2003 from *http://www.berghell.com/whitepapers/Permanent%20Digital%20Records%20and%20PDF%20Formats.pdf*

Kenney, Anne R., Oya Y. Rieger, and Richard Entlich (2003). *Moving theory into practice: digital imaging tutorial*. Retrieved March 1, 2003 from *http://www.library.cornell.edu/preservation/tutorial/index.html*

Lyman, Peter and Howard Besser (1998), Defining the problem of our vanishing memory: background, current status, models for resolution. In Margaret MacLean and Ben H. Davis, Eds., *Time and Bits: Managing Digital Continuity*. Los Angeles: The J. Paul Getty Trust.

NPES/AIIM (2002). Press release: "AIIM International and NPES partner to standardize use of PDF for document archive and preservation." Retrieved February 28, 2003 from *http://www.aiim.org/documents/standards/press_release-pdfa-aug02-2.pdf*

Ockerbloom, John Mark (2001, February 15). Archiving and Preserving PDF Files. *RLG DigiNews*, *5(1)*.

Teper, Thomas H. and Beth Kraemer (2002, January). Long-term Retention of Electronic Theses and Dissertations. *College and Research Libraries*, *63(1)*, 61-72.

Bibliography

ADOBE SYSTEMS INCORPORATED : PDF Reference, Third Edition: Adobe Portable Document Format Version 1.4. Boston: Addison-Wesley, 2001 partners.adobe.com/asn/developer/acrosdk/docs/filefmtspecs/PDFReference.pdf.

NEAL BEAGRIE AND MAGGIE JONES : The Preservation Management of Digital Material Handbook. 2002 www.dpconline.org/graphics/handbook/.

CCSDS SECRETARIAT : Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, BLUE BOOK. 2002 wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf.

STEVE GILHEANY : Permanent Digital Records and the PDF Format. A white paper based on a presentation at the ARMA 2000 conference. 2000 www.berghell.com/whitepapers/Permanent%20Digital%20Records%20and%20PDF%20Formats.pdf.

ANNE R. KENNEY, OYA Y. RIEGER, AND RICHARD ENTLICH : Moving theory into practice: digital imaging tutorial. www.library.cornell.edu/preservation/tutorial/index.html.

PETER LYMAN AND HOWARD BESSER : Defining the problem of our vanishing memory: background, current status, models for resolution . Hrsg.: Margaret MacLean and Ben H. Davis: Time and Bits: Managing Digital Continuity. Los Angeles: The J. Paul Getty Trust, 1998

NPES/AIIM : AIIM International and NPES partner to standardize use of PDF for document archive and preservation. 2002 www.aiim.org/documents/standards/press_release-pdfa-aug02-2.pdf.

JOHN MARK OCKERBLOOM : Archiving and Preserving PDF Files . RLG DigiNews. Band 5. 2001, February 15, 1, www.rlg.org/preserv/diginews/diginews5-1.html#feature2.

THOMAS H. TEPER AND BETH KRAEMER : Long-term Retention of Electronic Theses and Dissertations . College and Research Libraries. Band 63. 2002, January, 1, S. 61-72 ,