Long-term Retention of ETDs

Beth Kraemer
University of Kentucky

Tom Teper
University of Illinois at Urbana-Champaign

Overview

- Include slide with important summary points?
- ETDs are “born-digital” and archiving of born-digital documents is different.
- There are biases that come from our history of archiving paper that could negatively impact long-term retention of electronic documents.
- There are also biases based upon our relatively limited experience with managing digital information.
- Process choices are more important than specific format choices.

What is an ETD?

- ETD: Electronic Theses and Dissertations
- Producing TDs in wholly electronic format
- Impacts
  - Graduate School
  - Libraries
  - Students
  - Faculty

ETD Benefits

- Richer content possible
- Broader exposure of student work
- Master's Theses included, not just Doctoral Dissertations
- Faster availability of student work
- Full-text searching
- Student experience of preparing electronic documents valuable in the future workforce

ETD “Concerns”

- Will publishers publish material that is freely available on the web?
- Plagiarism easier with electronic originals
- Student must very strictly follow all fair-use and copyright standards
- Intermediate paper copies probably still necessary for Advisor and Committee to read
- Long-term electronic archival standards not yet established

ETD Programs

- Since 1994?? 199-something…
- How many institutions, etc.
- Most optional
- Most don’t require paper too
- Most maintain, or build upon, existing relationships with ProQuest (UMI)

ETD Programs

- Require the creation of databases to deliver documents via the web
- Require new systems for student submission
- Require training materials and/or programs
- Integration of Grad School and Library procedures

NDLTD

- International effort to promote/facilitate: NDLTD http://www.ndltd.org/
- Activities:
  - Talks
  - Publications
  - Annual Meetings
  - Listservs
  - Software
  - Federated Search
  - Documentation http://etdguide.org
  - Help!

NDLTD Objectives
(from NDLTD site)

- To improve graduate education by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing
- To increase the availability of student research for scholars and to preserve it electronically
- To lower the cost of submitting and handling theses and dissertations
- To empower students to convey a richer message through the use of multimedia and hypermedia technologies
- To empower universities to unlock their information resources
- To advance digital library technology

NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS

improving graduate education by developing accessible digital libraries of theses and dissertations

Objectives of NDLTD

- **To improve graduate education** by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing
- **To increase the availability of student research** for scholars and to preserve it electronically
- **To lower the cost** of submitting and handling theses and dissertations
- **To empower students** to convey a richer message through the use of multimedia and hypermedia technologies
- **To empower universities** to unlock their information resources
- **To advance digital library technology**

about NDLTD | browse/search ETDs | submit ETD | help/documentation
join NDLTD | community activities | research projects | related links

hot topics

- ETD 2002 Conference
- ETD 2003 Conference
- ETD Guide
- ETD Union Catalog

what's new on the website

- New NDLTD frontpage(s) [3 jan 2001]
  We welcome any feedback on the new design.
  Send your comments to etd@ndltd.org
- How to join the NDLTD Union Catalog [2 jan 2001]

 Comments? etd@ndltd.org Last modified: 3 January 2002

Example ETD

University of Kentucky – http://www.uky.edu/ETD

Masters with movie clips (mpegs):

THATS JUST THE WAY WE LIKE IT: THE CHILDRENS HORROR FILM IN THE 1980S

Author: Bentley, Christina Mitchell

Department: English

http://lib.uky.edu/ETD/ukyengl2002t00033/00cmbthe.pdf

mostly on the film's special effects and music, two components likely to draw in the other important audience for the film, older teenagers, especially boys.

Clip 1. Below, a playable film clip of *The Lost Boys* Trailer. Click on the still to play the trailer

Certainly, every film I have studied tried in some ways to address multiple audiences. While I have focused here on the films' attempts to draw in a child or young teen audience, it is important to note that all of these films engaged in various methods of drawing audiences, and that as the 1980s continued, and the "new" requirements for a successful film began to be more and more familiar, films became more savvy in appealing to multiple audiences. While *The Watcher in the Woods* and *Something Wicked This Way Comes* firmly established themselves as children's films through their association with Walt Disney productions, they also staked their hopes on aging Hollywood royalty (Bette Davis and Jason Robards, respectively) to appeal to parents and promoted their special effects in genre fan magazines like *Starlog* and *Cinefantastique*. *Gremlins*, which was also released in the first half of the decade, owes a great deal to the business sense of its producer, Stephen Spielberg, and was far more successful than either of Disney's entries into the genre.

Both *Gremlins* and *The Lost Boys* make significant efforts at cross-

ETDs are Electronic Documents

- Broader issues need consideration.
  - Generally require faculty and administrative understanding of issues.
- Similar to:
  - Electronic journal sites
  - Pre-print archives
  - Other university records that have never been paper
    - ETDs (like other university records) may have to comply with "permanent retention" requirements

Why is the Archiving of Electronic Documents a Hot Topic?

- Increase in “Born Digital” Documents
- Electronic is different from paper
- Future is unknown
- Recovery is more expensive than planning

Format of Original matters

- **Paper**
  - One format for all documents (paper!)
  - One format *within* document (solid paper!)
  - Eye-legible, “permanent” back-up
- **Mixed Media**
  - Paper with 5.25” floppies, Beta videotapes, etc…
- **Electronic**
  - Innumerable format types (PDF, XML, LaTeX…)
  - Can have multiple formats within one container
  - No innate back-up

Why Long-Term Retention is an issue for ETDs

- Prevalence of PDF as “archival” format for ETDs
- Lack of archival planning by ETD programs
  - Little information on this planning available on program websites.
- Unique nature of ETDs

Current Plans for ETD Retention

- Long term access to ETDs has not been tested.
- Plans include
  - Typical file back-up
  - Hardware upgrades
  - Migration of files to new versions or formats
- Almost can't be more specific that this, because we're planning migration to formats that don't exist yet!

Format requirements impact long-term access

- Format requirements
  - General discussion of XML vs. PDF
- Format recommendations
  - Recommend “archive friendly” formats (e.g., MPEG over AVI) and the use of developing standards such as METS (Metadata Encoding and Transmission Standard)
- Some programs accept whatever the student submits

Formats accepted by some NDLTD members

Institution	ARL	LG	CGS	Req.	Format
Louisiana State University	*	*	*	2002	PDF only
North Carolina State University	*	*	*	2002	PDF only
Pennsylvania State University	*	*	*		PDF only
University of Florida	*	*	*	2001	PDF only
University of Georgia	*	*	*		PDF only
University of Kentucky	*	*	*		PDF only
University of Tennessee, Knoxville	*	*	*		PDF only
Virginia Tech	*	*	*	1997	Prefer PDF, support others
West Virginia University		*	*	1998	PDF only
Western Michigan University		*	*		PDF only
Brigham Young University	*		*		PDF only
M.I.T	*		*		PDF only
University of Iowa	*		*		XML - suspended
University of Texas at Austin	*		*	2001	PDF only
University of Virginia	*		*		PDF or HTML
Vanderbilt University	*		*		PDF only

ARL – ARL Library
LG – Land Grant Institution
CGS – Council of Graduate Schools

Expectations about Archiving ETDs

- Electronic documents will have to be converted to be accessible and functional in the future.
  - Possible for document to have qualitative differences from the original!
- Archiving will be:
  - Active
  - Time-consuming?
  - Costly?
- “Benign neglect” = disaster!

What is “preservation”?

- Preservation – An umbrella term that concerns itself with providing access to materials for as long as needed by whomever needs them
  - Success is dependent upon cooperation.
- Asset Management – the business of providing access and protecting the institution’s investment.

Semantic Arguments I

- “Killing the Duck to Keep the Quack”
  http://www.acmi.net.au/FOD/FOD0055.html

  - by Simon Pockley
    - Short-term access
    - Long-term access
    - Preservation

Semantic Arguments II

- Digital Preservation – a singular term describing two activities.
  - Preservation through use of digital imagery – a possibility if done properly
  - Preserving digital information – a necessity considering the expense that goes into creating the information.

Requirements of "Digital Preservation"

- Make use possible
- Protect the original item
- Protect the surrogate
  - What about ETDs, when the surrogate is the original?

-Paul Conway, "Preservation in the Digital World"
http://www.clir.org/pubs/abstract/pub62.html

Requirements for Preserving Born-Digital Information

- Ability to be copied perfectly
- Ability to be accessed without geographic restraint
- Machine readability
- Ability to retain unique functionality.

-Jeff Rothenberg, " Avoiding Technological Quicksand"
http://www.clir.org/pubs/abstract/pub77.html

Preserving Electronic Records

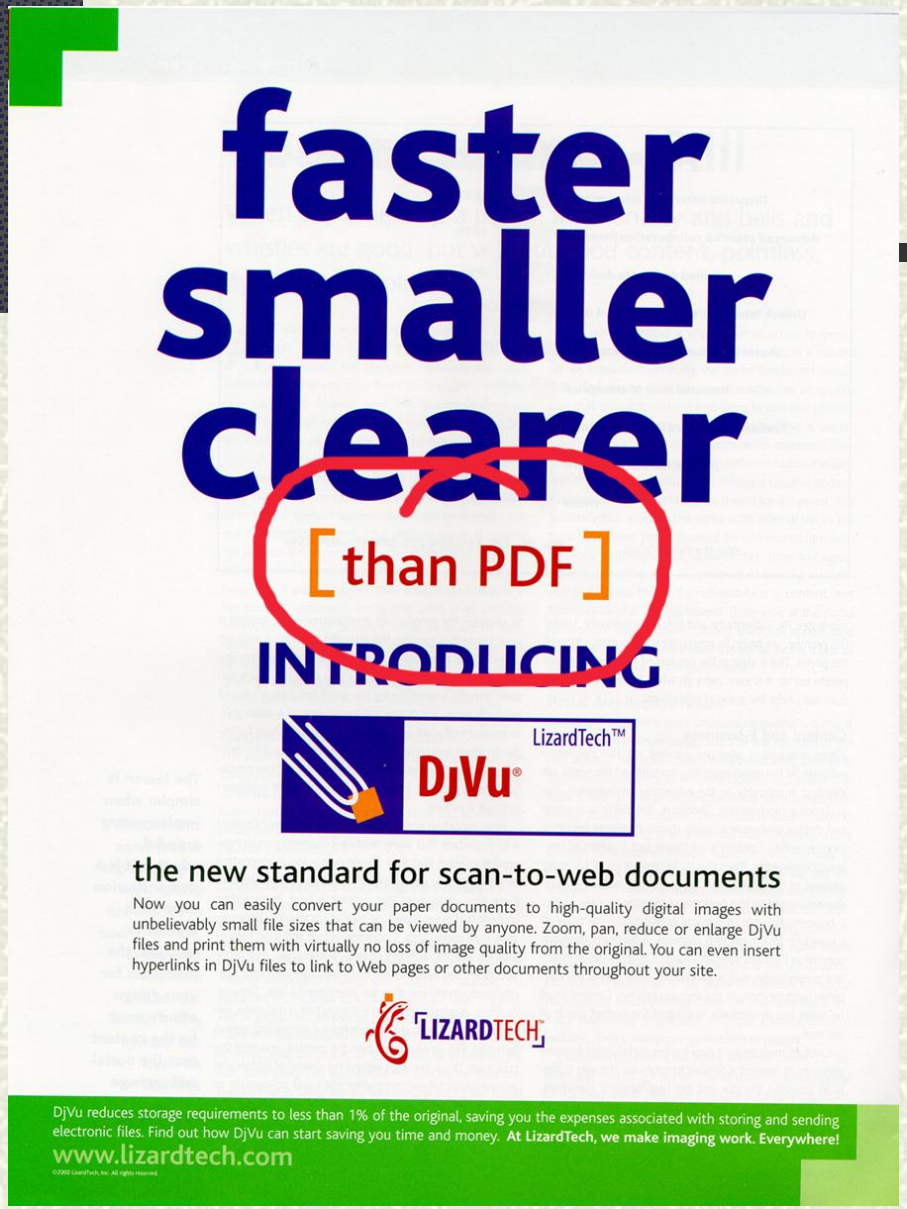
- Emulation – the process of using new technology to "emulate" the functionality of obsolete hardware/software on new technology
- Migration – according to CLIR and the RLG, a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to another.

Three Preservation Models

- Technology-based
  - LOCKSS www.rlg.org/events/pres-2000/reich.html
    **www.locks.org
  - CEDARS www.rlg.org/events/pres-2000/russell.html
  - NDIIPP www.digitalpreservation.gov/ndiipp/
- Analog
- Put it on the shelf and hope for the best

A Cautionary Tale…

- Who drives development?
  - The Government
  - Industry
    - Records Managers http://www.ArchiveBuilders.com
    - ProQuest (a.k.a., Bell & Howell or UMI)
    - DjVu, or having faith in non-standard



"Now you can easily convert your paper documents to high-quality digital images with unbelievably small file sizes that can be viewed by anyone. "

Why is PDF so popular?

- Ubiquitous delivery format
- Free reader
- Conversion from common word processors is fairly easy
- Submission format for many electronic journals
- Mimics paper
  - Page numbers always consistent
  - Layout always consistent
  - Prints well
  - Suitable for converting from paper - docs can become electronic but maintain paper feel

So what's wrong with PDF?

- Not meant to be an editor
  - Difficult to make changes
  - Easier to return to native format, edit, then re-create PDF
- **Migration to other formats is not easy**

Acrobat 5

- New features cite ability to “repurpose” PDF
  - As RTF (to get back to original Word)
  - As HTML
  - As XML
- Requires modification of Word from what you might normally use (recommended *before* you start work on your doc)
- Ability to extract is dependent on your original Word structure
- Resulting document can have missing elements (images), format errors (tables, line breaks)…

PDF summary

- Not push-button simple:
  - Can't just tell student "go make a PDF"
  - Maximizing the ability to export requires plug-ins and configuration that could be beyond some students
- Export options not very good
  - Moderately difficult
  - Resulting documents not 100% satisfactory
- Proprietary software
  - "Skills" in Acrobat not necessarily as transferable as mark-up skills, therefore, we are not really building anything educational into the process.

Factors Affecting the Ability to Archive Electronic Formats

- Well-documented
- Well-tested
- Non-proprietary
- Platform independent (hardware and software)
- If forced to choose proprietary, look at export formats

Short-Term Benefits v. Long-Term Costs I

- Preservation's motto: "Think Twice, Cut Once."
- Digital Projects: "Do it once, do it right"

  - Paul Conway – "Handbook for Digital Projects

Short-Term Benefits v. Long-Term Costs II

- Proactive Preservation v. Accident and Rescue
  - ISO Archiving Standards Working Group
  - National Archives of Australia

    -Seamus Ross, "Changing Trains at Wigan: Digital Preservation and the Future of Scholarship"
- More costly for born-digital items….

Short-Term Benefits v. Long-Term Costs III

- Memory

- The "Long Nineteenth and Short Twentieth Centuries"
  - Eric Hobsbawm, "The Age of Revolutions," "The Age of Industry," "The Age of Empire," and "The Age of Extremes"

The Future…

Two primary possibilities:

- Proprietary software formats will become archival standards
  - Generally lacks the features of archival standard
  - Technology tends to change!
- Popular formats will give way to other formats
  - Backfiles will have to be migrated or will be lost

…and Choices

- Scholarly Obligations
- Obligations of the ETD Community

What's an ETD program to do?

- Programs aren't necessarily including long-term access in their planning:
  - Attitude = "technology will find a way"
  - Attitude = "long-term access not important"
- Plans can't be specific in terms of format, but there are things you can do.

What's an ETD program to do?

1. Make Format Recommendations

- Argument from programs that accept anything: "limiting format limits creativity"
- There are formats that are more archive-friendly than others
- Migration headaches increase with multiple formats.

What's an ETD program to do?

2. PROCESS over FORMAT

- Particular formats will change and can't be predicted.
- Key process issues:
  - Maximize flexibility
    - non-proprietary formats
  - Review new formats regularly

What's an ETD program to do?

3. Stop viewing born-digital documents through paper vision

- One Document = One Format
- All document aspects deserve the same preservation attention (or lack…)
  - All parts (main body, text, appendix…)
  - All features (font, colors, page breaks…)

What's an ETD program to do?

4. Interact with ETD Authors

- We will be modifying their documents in the future
- Give EXPLICIT technical instructions
  - Can't just say "make a PDF"
  - Make creating an ETD an educational experience and a saleable skill.
- Accept new formats students choose for good reason

What's an ETD program to do?

5. Pay attention to developing standards in the preservation community.

- Monitor discussions on archiving issues
- Migration will be possible. How easy it is (cost effective, time efficient) depends on planning.
- Continue to work toward standards, but…
- Don't re-invent the wheel.

Selected Bibliography

- Thomas Teper and Beth Kraemer, "Long-term Retention of ETDs", *College & Research Libraries*, 2002, 63(1), 61-72.
- “File Formats to Support Long-Term Access to Electronic Records.” Managing Electronic Records 2000. Richard Fisher and Charles Dollar. 2000.
- Can Libraries of Digital Materials Last Forever?, Richard H. Ekman, Change, March/April 2000
- Best Practices for Digital Archiving: An Information Life Cycle Approach, Gail M. Hodge, D-Lib Magazine, January 2000.
- LC21: A Digital Strategy for the Library of Congress (2001), http://www.nap.edu/books/0309071445/html/
- Archiving the Digital Public Record: An Internet Snail’s Pace, Henry M. Gladney, iMP Magazine, Oct. 23, 2000.

Selected Bibliography

- Using Emulation to Preserve Digital Documents, Jeff Rothenberg, NELiNET, June 5, 2000.
- "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information." Council on Library and Information Resources and Research Libraries Group, ftp://ftp.rlg.org/pub/archtf/final-report.pdf, 1996.
- The Crisis in Electronic Government Record Keeping: A Strategy for Long-term Storage, Philip Coombs, Library Computing, 18(3), 2000.
- Phase II Report of the Working Group on Permanence of NLM Electronic Publications, 10/13/00, http://www.nlm.nih.gov/pubs/reports/permanence.pdf.

Selected Bibliography

- Ross, Seamus. “Changing Trains at Wigan: Digital Preservation and the Future of Scholarship.” London: National Preservation Office, 2000.
- Barnum, George D. “The Federal Depository Library Program Electronic Collection: Preserving a Tradition of Access to United States Government Information.” Presented at *Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials*. York, England: 2000. http://www.rlg.org/events/pres-2000/barnum.html.