

PRESERVING INTERACTIVE OR MULTI-VERSION
WEB-BASED DOCUMENTS

Rob Spindler, University Archivist

Arizona State University Libraries

rob.spindler@asu.edu

Third International Symposium on Electronic Theses and Dissertations

St. Petersburg, Florida, March, 2000

All rights reserved, Rob Spindler, 2000

Thank you for the opportunity to join you this morning. Today we're going to discuss the prospects for preservation of digital materials, with particular attention to web-based interactive or multi-version documents. As we get farther into this presentation, I believe we'll find it's necessary to have an understanding of some big picture goals before we can discuss specific digital preservation actions. As we'll see the nature of our goals directly impacts how our documents can be created and managed to increase the potential for their survival over time.

As a result I'd like to begin with the biggest question of all: What is a Dissertation?

- A "record" of research that documents fulfillment of one requirement for a doctorate, and is retained indefinitely OR
- A research publication kept only as long as the research is needed

[The audience voted on which definition they preferred and the vote was unanimously in favor of a "record of research".]

Since we've selected a "record of research" that needs to be kept indefinitely or for a very long period of time, we've set a very high bar for preservation of these materials. The answer to this question is important because there are two more critical decisions we need to make to effectively address preservation of any electronic document. They are:

- 1) How long must the document be kept?
- 2) Must a "record" of the document be kept?
 - Must the appearance and functionality of the document be retained Exactly, or Approximately?
 - Must we retain the context of the document as well as the document itself?

Duration of retention is an important preservation consideration in the context of our ever changing technologies. Our experience in the 1990's has shown that the interval between new releases of software and new operating systems is shrinking. In addition introduction of new storage devices, particularly proprietary backup systems, is also occurring more frequently over time. As the number of data storage format and software migrations necessary for long term retention increases, the potential for corruption or loss also increases.

For many years archivists have struggled with defining the difference between a file of data, a document, and a record. Most would agree that the essential difference between data and an electronic record is the context provided with that data, known as metadata.¹ There are however other more difficult considerations in defining a record. In the past we have generally thought of records as being produced in a fixed and tangible form, which is directly in conflict with our modern notion of web documents as living, changing virtual documents. Our ideas of what constitutes a record will need to be altered in order to accommodate the new possibilities of "born-digital" interactive documents.

After a number of years of receiving documents or data in obsolete formats with incomplete or absent contextual information archivists are realizing that we can't preserve just anything that comes to our door, and in fact most often we can't preserve electronic documents or files that are inactive and have not been used for a long time, or even a relatively short period of time. As a result we're beginning to think in terms of what can be done during the process of creation and during the active management phases of document life to increase the potential for survival of the document or a record of the document. In the context of electronic theses and dissertations these issues might best be addressed in the form of submission and/or format standards.

Let's assume we're going to keep a "record" of each electronic thesis or dissertation "permanently". Our next task would be to address six areas of concern:

SIX MAJOR CHALLENGES IN PRESERVATION OF ELECTRONIC DOCUMENTS

- Physical Obsolescence of Storage Media
- Physical Degradation of Storage Media
- Incompatibility/Non-Interoperability of Media
- Software/Operating System Incompatibility
- Metadata
- Human Error

The first three challenges relate to management of storage media, which may not be used for storage of active files or documents. Often this media is used to support backups, offline storage, or content distribution. Here are the media challenges followed by real or possible examples of each:

¹ For a very complete list of citations to information about metadata, see Anne Gilliland-Swetland's Metadata Presentation Abstract and Bibliography at <http://www.asu.edu/it/events/ecure/gilliland-swetland-present.html>. The leading metadata standard is known as the Dublin Core, and it is explained at the Dublin Core Metadata Initiative home page at <http://purl.org/dc/>. The most advanced research project specifically defining metadata elements for electronic records of universities is in progress at Indiana University. See Indiana University Electronic Records Project, *Metadata Specifications*, 1998. <http://www.indiana.edu/~libarche/metadataspecifications.html>

- Physical Obsolescence of Storage Media

“5.25 inch floppy doesn't fit in my CD-R drive”

- Physical Degradation of Storage Media

“My CD doesn't read anymore since the recording surface has been degraded by air pollutants.”

- Incompatibility/Non-Interoperability of Media

“My Phillips DVD won't play in the fifty Sony DVD players we purchased for the library.”

In general the storage media challenges seem to have received a great deal of attention in the popular media and the professional literature, but in my opinion they are the easiest to manage. There is some very good science out there on the reliable shelf life of various media and interoperability of certain media formats like CD-R²; We can project the expected physical life of the media, track changing format standards for drives and plan to migrate the data to new media based on the expected physical shelf life and availability of new formats. The generally accepted reliable shelf life of CD's for example is between 5-30 years, depending on manufacturer, storage conditions, level of use, etc. However, as we heard from Delphine Lewis yesterday UMI has now converted to refreshing CD's every three years. Delphine says this is mainly because the physical obsolescence and incompatibility/non-interoperability pieces are changing faster than it takes for the media to physically degrade.

Regardless of the issues in media refreshing and migration, I believe the most difficult challenges in digital preservation reside in the last three areas, also given below with illustrative examples:

² National Technology Alliance, *Doculabs Test Report: Compatibility of CD-R Media, Readers and Writers*, Chicago, IL, Doculabs Inc., November 30, 1996 (amended June 30, 1997). Formerly available at <http://www.nta.org>. Studies on media stability for various electronic storage media developed by Dr. John VanBogart of the National Media Laboratory were formerly available through the National Technology Alliance site, but have since been removed. The findings were removed from the web as a result of inaccurate press reporting and erroneous interpretation by the public that distorted the specifics of the research, according to Dr. VanBogart.

- Software/Encoding/Browser Incompatibility
 - New software release won't run documents from old release
 - New software release opens old file but contents are corrupted
 - Product designed with Netscape doesn't look the same in Internet Explorer
 - Proprietary codes from HTML editing packages do not convert to XML

In this area we are faced with documents that won't open at all or documents that are corrupted and must be repaired if possible (usually through expensive hand editing and correction of content or comprehensive manual re-entry). The two major strategies for addressing software issues are known as emulation and migration. The emulation concept proposed by Jeff Rothenberg of the Rand Corporation involves writing new software that mimics the appearance and functionality of old software. Recently experts in this area have challenged Rothenberg's work, suggesting that efforts to write emulation software solutions have not resulted in exact reproductions of appearance or function.³

The migration strategy tends to be the more widely accepted option at this point, but we are still seeing examples of unsuccessful migrations to new releases of the same products, or corruption resulting from migrations across different software packages. Another important question is the longevity of the software vendor and related support for their products. Many document managers are selecting ubiquitous software packages, relying upon a large "installed base" in the hope that that the manufacturer will survive for many years and that future releases will have complete backwards compatibility. It's my sense that the larger software vendors are paying more attention to these issues and increasingly building backwards compatibility and limited interoperability functions into their products. As a result we seem to be making progress on these issues, but for the time being we can still expect mixed results until market forces demand complete compatibility and interoperability from our vendors.

3. Rothenberg, J. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Washington, D.C.: Council on Library and Information Resources, 1999.

<http://www.clir.org/pubs/reports/rothenberg/contents.html> Clifford Lynch expressed his doubts about emulation strategies in a presentation at Arizona State University on February 15th, 1999, citing early unsuccessful efforts to emulate simple video games like Pacman. David Bearman directly addressed Rothenberg's findings in "Reality and Chimeras in the Preservation of Electronic Records" ³ *D-Lib Magazine*, 5(4): April 1999. <http://www.dlib.org/dlib/april99/bearman/04bearman.html>

The minor differences in appearance that occur when the same object is opened in different browsers may not seem to be a problem, but if we intend to preserve the appearance and functionality of a “record”, browser compatibility becomes an issue. Textual website content is at less risk in this scenario, but as we move into multimedia presentations and more substantial and meaningful graphic or video content these features may become essential elements of the record that need to be preserved. There are similar issues related to use of proprietary HTML codes that do not meet the W3c standard tag sets, in that it's probably not that difficult to change out the offending codes so the documents parse in XML, but what happens to the appearance and functionality?

- Descriptive and Administrative Metadata

Metadata is the key to retaining the context of a given electronic document or record. As Cliff Lynch of the Coalition for Networked Information said at my university in 1999 “We know a lot about retaining bits and bytes but often we can't make sense of them when we have them...”⁴ It seems that development of descriptive metadata standards such as the Dublin Core and the Resource Discovery Format have largely been driven by internet commerce or private sector concerns, and there is good progress in developing these standards to facilitate automated search and retrieval of content.

The place where I believe we need to pay more attention is in the area of administrative metadata, which should address issues of:

- Version control
- Software used/required
- Location/availability of data dictionary/software documentation
- Location/availability of backups
- Migration history – When/how was the document last migrated

Those of us with backgrounds in computing remember the data dictionary or code book approach to documenting databases and flat files. What does the data dictionary for an interactive website look like? Where is it stored? How is it read or delivered for use? The migration history part has a nice analogy in health care – when we consider treating a person's illness we generally look at their health history. If we have a corrupted, damaged or just forgotten document it's very useful to know where it has been and how it has been managed in order to facilitate repair or possible replacement of the file from an accurate backup. It seems we haven't really come very far in deciding how and where this kind of information should be captured.

⁴ Clifford Lynch presentation at Arizona State University, February 15th, 1999.

The final frontier in digital preservation is one I've incorrectly entitled Human Error, since some of our challenges are intentional as well as accidental or erroneous actions. Here are some examples:

- Accidental deletion
- Website is vandalized
- Backups not performed properly (wrong files backed up)
- Backups not performed properly (Video server backed up, web server not)
- Backups not performed properly (Backups of different servers not contemporaneous)

The backups issues are key ones in the web environment since the timing and scope of backups are critical for capturing a "record" of an interactive website. Many of our more complex web products are now stored in a number of different servers to facilitate efficient retrieval and display. The products themselves only exist for a moment in time on a patron's computer screen, but the sources of the content are distributed. On one level capturing a record of the content requires that all the sources of product components are backed up so that the record of the website is not missing certain pieces. On the other hand the institution may consciously decide that a student chat room associated with a given product is not part of the record and should not be retained and made available for privacy reasons. We need to evaluate all the potential components of the record so we may judiciously select those components that meet our institutional recordkeeping needs.

There is a similar concern with the timing of backups. In an interactive website certain components of the site may be static and others changing. Ideally the backup or capture of the record components across a series of servers needs to be done contemporaneously so an accurate snapshot of the components and their relationships can be retained. In addition snapshot timing needs to be established with some attention to the rate and nature of change in the website. If real time updating of the website is constantly in progress, when should the snapshot be taken? At Arizona State we have limited our updating of web-based university policy manuals to three update periods a year so that we can capture all the changes to the site and document the dates of implementation for new policies.⁵

⁵ See Robert P. Spindler, "Preserving Web-Based Records", Preservation and Access for Electronic College and University Records Conference Presentations, Arizona State University, December, 1999.
<http://www.asu.edu/it/events/ecure/spindler-presentation.html>

CASE STUDY

In 1995 Keith Voegele, a doctoral candidate in the Computer Science Department at Arizona State University, received approval of his web-based interactive dissertation entitled *Tessellation of Bibliographic Data: An Example Using Categorical Data*. The dissertation consisted of three components, a website containing citations to literature concerning web-based visualization technologies and some text describing creation of the site; a recordable CD where the student "archived" the C and HTML files that comprise the site (at the request of his review committee); and a hardcopy volume that contained a bibliography of the citations to visualization literature and some text about creation of the site and how it worked. The website included an interactive "tessellation" diagram, in which site visitors were invited to add their own new citations which would then be automatically plotted in an electronic diagram. In my opinion the most complete version of the dissertation was the website itself.

The student's committee demonstrated remarkable foresight in requiring the storage of site files on CD-R, but the hardcopy documentation did not convey when the CD-R had been written and whether the review committee saw the same version of the site that was copied to CD-R or something else. Since the site was interactive we had lost the opportunity to preserve an exact copy of the document that was approved by the committee. When we attempted to open the CD-R we discovered it was formatted for Macintosh computers (we had only one Mac in the University Libraries) and the files were not write protected in any way.

In an email exchange and conversation with the author we discussed the potential for rebuilding the site from the hardcopy documentation and the files on CD-R. Voegele suggested it would be impossible to recreate the site from these sources since there were certain compilers and other pieces of software resident on the server which could not be copied to the CD-R.⁶ As a result we could probably display some pages that looked like the pages in the site, but could not create an exact or even a near reproduction of its functionality. In March 2000 I returned to the site for the first time in a few years only to discover it had been deleted from the College of Engineering server.

⁶ Electronic mail from Keith Voegele to Robert Spindler, October 23, 1995.

RECOMMENDATIONS:

Given the variety of applications and circumstances it's very difficult to make specific recommendations, but there are some general practices that can be followed to increase the potential for survival of electronic theses and dissertations:

- Determine your recordkeeping goals for accuracy and length of retention.
- Identify the components to be preserved and the environments used in the document.
- Select and label robust storage media.
- Establish a media refreshing and migration plan.
- Carefully select and regularly upgrade backup systems.
- Design your product for use with a specific browser.
- Use widely-implemented software.
- Document your browser and software selections.
- Avoid proprietary document encoding/use W3c verification.
- Migrate your legacy documents along with your current documents.

CONCLUSION:

This high standard for retention of ETD's is not intended to be applied to any website since many websites are not "permanent" archival records. Rather I believe the importance of theses and dissertations as a student record and as a valuable research source demands long term and highly accurate retention. Others may disagree, but it is essential that each institution define their goals so appropriate strategies can be employed to maximize the potential for survival of the documents at the level of rigor that meets your institutional needs.

The biggest hurdles in electronic preservation are the human factors, particularly resistance to organization and planning, but forethought is necessary to increase the odds in favor of survival in this rapidly changing technological landscape. The question we all face in digital preservation is addressing the need to adapt the process of creation without sacrificing the beauty of creative endeavor. We will reach this balance only when we have established effective partnerships between creators of documents, and those charged with stewardship of those products. In the meantime, important artifacts of our scholarly productivity and communication will be left at risk, leaving our children to wonder why so much of their parents legacies has been lost in the ever shifting sands of time.