9th International Symposium on Electronic Theses and Dissertations
IXe Symposium international sur les thèses et mémoires électroniques
June 7 – 10 Juin, 2006, Quebec City / Québec, Canada

DAITSS: ANOTHER PRESERVATION OPTION FOR ELECTRONIC THESES & DISSERTATIONS

Priscilla Caplan
Florida Center for Library Automation
pcaplan@ufl.edu

Chuck Thomas
Florida Center for Library Automation
cthomas@ufl.edu

1. ABSTRACT

Electronic theses and dissertation (ETDs) are an important and valuable part of the growing body of digital materials that must be preserved by college and university libraries. This paper reviews some of the current options and applications available to help implement bona fide digital preservation environments for ETDs. DAITSS, a new product being developed by the Florida Center for Library Automation for this purpose, is explained and differentiated from other available tools.

2. CURRENT OPTIONS

As an increasing number of institutions are moving to digital theses and dissertations, the need to address the long-term preservation of ETDs in digital form becomes more pressing. Preservation is more than establishing bibliographic control, ensuring secure storage, and encouraging the use of "preservation-friendly" formats. While these are necessary steps, preservation planning should also include the implementation of active preservation strategies.

There are several current options for the long-term preservation of ETDs. One option is to rely upon a paper or microfilm copy of the ETD as the permanent archival copy. This has the advantage that we know how to handle these media and have some confidence in their preservation qualities. However, the process of deriving a paper or microfilm preservation master risks losing functionality not supported by these media, such as internal or external linkages and embedded or associated media objects like video and audio files.

A second approach is for the graduate school to require ETDs to be created in certain formats considered to be more "archivable" than others. Many schools now require theses to be submitted in PDF or HTML, some encourage PDF/A, and some are considering or supporting XML encoding. Supplementary files in other formats may be accepted, but they are considered add-ons rather than integral parts of the dissertation. It is wise to be concerned about the preservation properties of digital formats, and probably good policy to set guidelines for what can be submitted. However, this alone does not constitute a concrete preservation plan for managing the preferred format.

A third approach is to contract with commercial third parties that offer preservation services. "Bit level" preservation services guarantee digital files are unaltered and readable from media though methods such as secure storage, redundancy, and fixity checking. "Full" preservation services include strategies to counter format obsolescence, such as forward migration, normalization, and/or emulation.

The OCLC Digital Archive was one of the first third-party options available. Although only bit-level preservation is supported at this time OCLC intends to move to full preservation treatment for harvestable

digital objects. The OCLC Digital Archive is available for a wide variety of formats and content, and services can be customized depending upon customer needs (OCLC, 2006). Their goal is to eventually meet the certification criteria of a Trusted Digital Repository (RLG, 2005).

Another player in the emerging third-party market is ProQuest/UMI, a long-time service provider for libraries and universities. Since 2002, ProQuest has teamed with the Berkeley Electronic Press in a collaborative Digital Commons institutional repository initiative for online access to digital research papers such as ETDs. ProQuest also now markets a Digital Archiving and Access Program. Like OCLC, the ProQuest archive plans to move beyond bit-level services to full preservation, and aims to become a certified as a trusted digital repository (McLean, c.2004).

A fourth option for archiving ETDs is for an institution to manage its own preservation repository. D-Space, an open-source institutional repository system from MIT, is a popular platform that accepts all kinds of digital content and can accommodate ETDs. DSpace now provides bit-level preservation, but mechanisms for full preservation strategies are being developed by the open source community under the auspices of the DSpace Federation. Several other open source products compete with D-Space, including ETD-db from Virginia Tech and the popular E-Prints software from the University of Southampton. However, many of these tools are primarily designed for access and use-copy management, and would have to be combined with other applications and processes to constitute a true archiving environment Fedora, another open-source product, provides great promise for organizations able to devote resources to building onto its underlying storage and dissemination architecture. Some institutions, such as theUniversity of California, are building their own state-of-the-art digital repositories, or mixing a variety of applications together to provide the services and processes of a true preservation environment as described in the OAIS framework. Finally, commercial systems such as DigiTool from Ex Libris are finally coming into the library systems market with built-in tools enabling better digital preservation management.

3. THE DAITSS OPTION

The Florida Center for Library Automation (FCLA) is pursuing this fourth option on behalf of its partners. FCLA is a state agency created to support the automation needs of the libraries of the public university system of Florida. FCLA runs the Florida Digital Archive (FDA) as a preservation repository for the collective use of the eleven state universities. The primary motivation for building the Florida Digital Archive was the desire of the library directors to assure themselves of the long-term preservation and usability of the ETDs for which they were responsible. The libraries can batch-submit ETDs and other materials to the repository for archiving, and they can request archived materials back in usable form at any time. The FDA began operating in November 2005, and as of June 1 2006 it has ingested 11,786 information packages (digital books, photographs and ETDs) comprising 118,918 files requiring 2.7 TB of storage.

The software application underlying the FDA is named DAITSS (Dark Archive In The Sunshine State). DAITSS is a Java application that runs under Linux and uses MySQL for the management database (FCLA,2005). It is designed as an implementation of the major functional areas of the Open Archival Information System (OAIS) framework: Ingest, Data Management, Access, Archival Storage, Preservation Planning and Administration. It uses established and emerging library standards wherever possible, including the PREMIS core preservation metadata element set, Z39.87 technical metadata for digital still images, and METS.

DAITSS is a system for digital preservation only; it has no other function. Unlike some repository systems initially designed for storage and access and now being revised to support preservation, DAITSS was designed from the beginning to implement the active preservation strategies of format normalization, mass migration, and migration on request. In order to perform full preservation for a format, Java classes to handle the file format and any relevant bitstream formats must be programmed and included in the DAITSS libraries. DAITSS is designed to be extensible to any number of formats.

All preservation strategies are implemented in the Ingest function as materials enter the system. Every incoming file is preserved in perpetuity exactly as it is received, but certain other versions may be created depending on the file format. A normalized version will be created if the incoming format can be transformed into a more "preservable" version. For example, a CinePak video stream within an incoming Quicktime file will be normalized to Motion JPEG. A localized version will be created if the file contains external links that can be downloaded, stored locally, and referenced by a relative file path. For example, an XML file that references a schema and a stylesheet will be localized if possible to facilitate future validation. A migrated version will be created if the file format is considered in danger of obsolescence and a successor format exists. Content stored in the repository can never be changed, but it can be disseminated and re-ingested if a format migration is necessary in the future.

DAITSS is a "dark archive" in that it supports no online real-time public access. Archived materials can be accessed only through the dissemination function, which allows authorized users to request copies to be delivered asynchronously. The standard dissemination package consists of the content as originally submitted along with the latest "best" version of any file, if different.

Although DAITSS can ingest any type of material, it is particularly well suited to ETDs because it is totally independent of any submission and access systems implemented by the university. Schools can use their own cataloging, search, access and presentation services for their online ETDs, or they can rely on ProQuest/UMI's digital collection. For preservation, they would send a copy of the ETD file(s) along with a METS-format descriptor to DAITSS. If for any reason the online copy was threatened by the obsolescence of a file format, the school would request dissemination of the current version of the master ETD from DAITSS. This would be used to replace or re-generate the online version.

FCLA has a commitment to the Institute of Museum and Library Services (IMLS), which partially funded DAITSS development, to make the application widely known and freely available. We anticipate releasing the system under an open source license for implementation and broader co-development in calendar 2005. The system is designed to make it easy to distribute the development of format classes, which means that as the number of DAITSS implementations and developers increases, the number of formats it will be able to handle for full preservation should increase proportionately.

Although other options are available as described above, FCLA believes the marketplace for digital preservation services and tools benefits from the additional choice of using DAITSS. FCLA wants to make DAITSS known to the international community and ETD institutions as an option in their preservation planning. Libraries who might be interested in pursuing a development partnership arrangement should contact Priscilla Caplan at pcaplan@ufl.edu or 352-392-9020 x324.

4. REFERENCES

FCLA. (2005). "DAITSS Overview." Accessed May 28, 2006 online at http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf.

Austin McLean, "Dissertation Archiving and Access: A Case Study for Accessibility and Preservation." Accessed May 28, 2006 online at http://www.il.proquest.com/umi/temppages/daap/DAAP_whitepaper.pdf

OCLC (2005). "Digital archive preservation policy and supporting documentation." (20 January revision). Accessed May 28, 2006 online at http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf

RLG. (2005). "An audit checklist for the certification of trusted digital repositories." (August, draft for public comment). Accessed May 28, 2006 online at http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf