ARTICLE EXCERPTED FROM:

ISQ INFORMATION STANDARDS QUARTERLY

SPRING 2010 | VOL 22 | ISSUE 2 | ISSN 1041-0031

SPECIAL ISSUE: DIGITAL PRESERVATION

DIGITAL PRESERVATION METADATA STANDARDS

TRUSTWORTHY DIGITAL REPOSITORIES

UNIFIED DIGITAL FORMATS REGISTRY

AUDIO-VISUAL DIGITIZATION GUIDELINES

DIGITAL PRESERVATION EDUCATION

NISO
How the information world CONNECTS

Melissa Goertzen

[CONFERENCE REPORT]

MELISSA GOERTZEN

What It Takes to Make It Last: E-Resources Preservation: A NISO Webinar

Since pen has been put to paper, memory institutions have been tasked with the responsibility of preserving cultural and intellectual heritage. For centuries, experts have created storage environments that protect tangible materials from decay, while at the same time allowing visitors to access information. Now, thirty years into the digital revolution, it has become clear that preservation policies must expand to include digital content if information is to remain available for future generations. Factors such as media obsolescence, degradation, and server failures have left virtual content in a vulnerable position, and sustainable platforms of preservation must be utilized to prevent information from vanishing without a trace as technology trends shift.

In the past, the preservation of physical collections fell on the shoulders of institutions. It is becoming increasingly apparent that due to the extensive and interconnected nature of the digital universe, electronic preservation efforts must be addressed on a communal instead of institutional level. As a result, experts are engaging in collaborative projects and public discussions to address sustainability challenges presented by digital content. On February 10, 2010, NISO hosted a webinar entitled *What it Takes to Make it Last: E-Resources Preservation*, which addressed topics relating to digital memory and preservation repositories at academic institutions. Speakers included **Priscilla Caplan**, Assistant Director for Digital Library Services at the Florida Center for Library Automation, and **Jeremy York**, Assistant Librarian at the University of Michigan Library. Each presentation provided examples of preservation standards and policies that can be implemented by universities to ensure digital content will remain accessible for years to come. The session was both informative and interesting, and provided participants with ideas regarding how to approach issues of content sustainability.

The session was both informative and interesting, and provided participants with ideas regarding how to approach issues of content sustainability.

The webinar began by challenging some preconceived notions regarding preservation. At times, it seems that preservation efforts come down to storing duplicate copies of digital content on external hard drives or DVDs, essentially creating the equivalent of electronic photocopies. As **Priscilla Caplan** explained, content availability is only one piece of the puzzle. In order for archival structures to be successful, they must also guarantee usability. This includes ensuring that the quality is not altered, the files are displayable, and that each object remains what it proposes to be. In other words, preservation is not simply about capturing snapshots of the original. It is about ensuring that information remains available in its authentic form for future generations.

Because the preservation of electronic content is a complex issue, it can be challenging to know where to begin. For example, what procedures or policies do institutions need to consider to ensure that digital output is ready to be ingested into communal preservation repositories? Caplan explained that standardization is the key, and that metadata and object files can be prepared for long-term preservation by following established guidelines and framework models. One such source is the *Open Archival Information System (OAIS) Reference Model*, which creates a basis for standardization by providing a common vocabulary that can be used to describe and compare

CONTINUED »

Ten years from now will it be possible to access and edit the many PDF files that have been burned onto DVDs for archival purposes? Because it is impossible to anticipate how technology will evolve over the next several years, Caplan suggested the use of sustainable formats wherever possible.

archives. Through this framework, institutions can discuss how preservation policies and procedures may change over time as technology evolves, and what information is needed to ensure the future accessibility of materials within a designated community of users.

In addition to guidance found through archival reference models, institutions can consult checklists or request audits to ensure they have established trustworthy repositories. Processes such as the Trustworthy Repositories Audit & Certification (TRAC) or the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) can assess the reliability and readiness of organizations to take on the responsibilities of long-term preservation. Components that are evaluated include organizational infrastructure, technical infrastructure, digital object management, documentation, and transparency, just to name a few. Such checklists can be of great value, as they can alert system administrators to issues that may put content at risk if left unchecked, such as media obsolescence and degradation.

While it is necessary to back up information using media devices, it is vital to ensure that information stored today will be accessible tomorrow. For example, ten years from now will it be possible to access and edit the many PDF files that have been burned onto DVDs for archival purposes? Because it is impossible to anticipate how technology will evolve over the next several years, Caplan suggested the use of sustainable formats wherever possible. For example, institutions can opt to use PDF/A formats instead of PDF files. The difference between these formats is that the former produces self-contained files that embed information such as font, color, and preservation metadata into the document code. As a result, experts suggest that PDF/A files will maintain their integrity and allow for authentic reproductions to be created for years to come.

Caplan began her second presentation on the topic of preservation metadata, by introducing the *PREMIS Data Dictionary*. She explained that PREMIS supports preservation metadata by requiring information relating to content viability, renderability, fixity, and authenticity. This ensures that material remains readable and usable, and that each source is verifiable and without alteration. Components required by PREMIS include characteristics of administrative, technical, and structural metadata, and consideration is given to maintaining relationships that exist between objects and records stored in repositories. As Caplan explained, the goal of preservation metadata is to include all information that a repository requires to ensure long-term sustainability. To this end, PREMIS is designed to define required information, create and manage collection records, and prepare metadata for entry and use in automated workflows. It supports the creation of core metadata, which in turn provides repositories with all information required to support long-term preservation efforts and work towards content sustainability.

When developing preservation metadata and preparing material for use in repositories, Caplan provided several tips that should be kept in mind. First of all, institutions should standardize metadata using data dictionaries in order to ensure that all necessary information required for preservation is in place. It is also important to test standardized metadata records to guarantee the success of imports and exports. Finally, all information about creation applications and environments should be recorded and embedded into the document code if at all possible. When these elements are present, institutions are ready to take on the responsibility of long-term content preservation.

Following Caplan's presentations, **Jeremy York** spoke about his involvement with the HathiTrust Digital Library project. His presentation provided webinar participants with

an excellent example of how collaboration, preservation audits, and the creation of preservation metadata work together to create stable and sustainable preservation repositories. The HathiTrust Digital Library was launched by the University of Michigan in 2008, and has since grown to include 4.6 million volumes. The initial focus of the project was on digitized book and journal content, but has expanded to include born digital content as well. The initiative began as a collaboration between thirteen universities on the University of California system and the Committee on Institutional Cooperation. In practice, HathiTrust was set up to provide a stable repository in which institutions could contribute digital content for long-term preservation. The central goal of the partnership was to preserve human knowledge for the common good, and provide continuous access to research communities.

As York explained, one of the strengths of HathiTrust is that it relies on a combination of expert staff and input from preservation communities to ensure content remains available for future generations. Partners are provided with the opportunity to communicate with project managers, information technologists, and copyright officers who can provide suitable information regarding servers, content migration, and storage. In addition, the repository is built around standardized procedures outlined in the *OAIS Reference Model,* and documents all preservation actions in accordance with standards outlined by the *PREMIS Data Dictionary.* Recently, HathiTrust also underwent several system audits to ensure the repository is in a position to take on the responsibilities of long-term preservation. For example, the project was reviewed by the Digital Curation Center and Digital Preservation Europe using standards outlined by DRAMBORA. The working elements of collaboration, standardization, and the use of repository audits has allowed HathiTrust to guarantee content sustainability and provide valuable information to an international community of researchers.

There are several ways in which HathiTrust strives to maintain the authenticity, integrity, and usability of its contents. First of all, the repository is structured to preserve the layout and general appearance of content deposited in the repository. It focuses on the development of preservation metadata, and records information pertaining to the creation of content deposited in the digital library. Also, only sustainable formats are ingested into the repository. These include TIFF, JPEG or JPEG2000 files, all of which can be easily migrated over time. Data integrity is maintained through means such as checksum validation. Finally, to guarantee efficient storage of files and protection against system failures, HathiTrust has set up two clustered storage systems located in separate geographic locations. Also, a third encrypted backup is stored in a facility at Ann Arbor. The use of preservation metadata, sustainable file formats, and off-site servers are only several of the ways the HathiTrust Digital Library works with partners to ensure that digital content is highly accessible to the research community, and will remain available for years to come.

The information provided by both Priscilla Caplan and Jeremy York served to provide excellent overviews of the challenges and rewards of digital preservation. Webinar participants were provided with many useful suggestions and resources that can be implemented when designing digital collections or building preservation repositories. Through the efforts of organization such as PREMIS and the HathiTrust Digital Library, memory institutions have great hope of providing continuous access to authentic digital content for generations to come. | CR | doi: 10.3789/isqv22n2.2010.12

MELISSA GOERTZEN <mjgoertz@ucalgary.ca> is Project Manager in the Digitization Unit, Libraries and Cultural Resources, at the University of Calgary.

RELEVANT LINKS

NISO E-Resources Preservation Webinar Slides
www.niso.org/news/events/2010/preservation

Digital Curation Centre, Digital Repository Audit Method Based On Risk Assessment (DRAMBORA).
www.dcc.ac.uk/resources/tools-and-applications/drambora

Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1), ISO 19005-1:2005
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

HathiTrust Digital Repository
www.hathitrust.org

PREMIS Data Dictionary for Preservation Metadata
www.loc.gov/standards/premis/

Space data and information transfer systems – Open archival information system Reference model, ISO 14721:2003
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

Trustworthy Repositories Audit and Certification (TRAC) criteria and checklist, Chicago: Center for Research Libraries; Dublin, Ohio: OCLC Online Computer Library Center, Inc., 2007.
www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying