ISQ

INFORMATION STANDARDS QUARTERLY

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

CR

Priscilla
Caplan

PRISCILLA CAPLAN

NIST Digital Preservation Interoperability Framework

The National Institute of Standards and Technology held a workshop on developing a roadmap for a Digital Preservation Interoperability Framework on March 29-31 on the NIST campus in Gaithersburg, MD. The purpose was to identify U.S. requirements, technologies, and best practices for standardization related to long-term preservation. A second workshop was held on April 21–23, 2010, in Dresden, Germany. Results of the two workshops will inform the efforts of the ISO/IEC JTC1 Study Group on Digital Content Management and Protection (SGDCMP), which was reconstituted in 2009 with a specific focus on digital preservation.

The U.S. workshop, which attracted about 130 registered attendees, was organized into three tracks: content, technology, and standards. Each track occupied a day of the meeting, beginning with a keynote address and ending with a panel discussion. In between, speakers were allotted 30 minute slots to present papers they had previously submitted. The format included time for questions after each talk and as part of the panel presentations, allowing ample opportunity for audience participation.

➡ The whole-conference keynote was delivered by **Chris Greer**, the Assistant Director for Information Technology Research and Development in the White House Office of Science and Technology Policy. Dr. Greer focused on challenges to the preservation of scientific data, including the great diversity in patterns of information use and exchange in different disciplines, the need for data management expertise and infrastructure, the need to incentivize data management planning, and the need for sustainable economic models for preservation and access.

Greer's talk set a good tone for the workshop, which presented a better mix of attendees from scientific and cultural heritage domains than the typical preservation conference. Government agencies were well represented, with speakers from the NSF, OSTI, NASA, USGS, NOAA, and NIH as well as NARA, the Library of Congress, the Smithsonian, and the Government Printing Office. The library domain was represented by the usual suspects presenting on PREMIS, the MetaArchive, DigCCurr, LOCKSS, and DAITSS.

Although all of the presentations were interesting, only a minority concerned either standards or interoperability.

➡ **Dr. Walter Warnick** from the U.S. Department of Energy, Office of Scientific and Technical Information (OSTI) gave a talk entitled *The Interoperability Solution: Federated Search and Good Databases*. Noting that search engines have trouble indexing the deep web where much scientific data resides, he touted the success of Science.gov and WorldWideScience.org, federated search gateways which transform queries into target-specific searches. While it is nice to see federated search actually working, the talk did not address digital preservation directly.

> The workshop presented a better mix of attendees from scientific and cultural heritage domains than the typical preservation conference.

➡ **David Minor** spoke about a project to develop tools and methods to automate the exchange of data between two approaches to preservation storage: the MetaArchive Cooperative and Chronopolis. The MetaArchive is a Private LOCKSS Network while Chronopolis is a federated data grid

HDF5 is an open source file format for storing and managing data, and its associated tools and applications. The HDF5 technology suite is used extensively by NASA and other agencies dealing with extremely large datasets.

based on iRODS (Integrated Rule-Oriented Data System). The project plans to examine the atomic units in each system's processing (ingest, verification, data transfer, fixity) to identify commonalities and differences and develop an ingest reference model. It will also draft a standard XML representation of common technical metadata that needs to be tracked.

➡ **Joseph Pawletko** from New York University described TIPR (Towards Interoperable Preservation Repositories), another system-to-system interoperability project. In contrast to the MetaArchive/Chronopolis project, TIPR is oriented towards OAIS-based repositories and assumes that digital provenance and rights information must be maintained across any package transfer. TIPR has defined a common transfer format called the RXP (Repository Exchange Package), based on METS and PREMIS. The RXP may be a candidate for further standardization activity.

➡ **Leslie Johnston** from the Library of Congress (LC) presented on BagIt, another specification for packaging digital content for transfer. BagIt is a simple format consisting of a bag of content and a simple manifest, and in fact is used by the TIPR project to carry RXPs. LC has developed a number of BagIt tools, including a validation script, a verification script for fixity checking, a parallel retriever script for efficient package transfer, an authoring tool, and others.

➡ **Dr. Mike Folk** spoke about HDF5, which is an open source file format for storing and managing data, and its associated tools and applications. The HDF5 technology suite is used

extensively by NASA and other agencies dealing with extremely large datasets. It not only enables the management and manipulation of this data, but is a good archivable format for long-term preservation. **Matthew Dougherty**, a researcher at the National Center for Macromolecular Imaging, spoke about the proliferation of proprietary file formats in his field. High resolution electron microscopy has ten formats, while optical x-ray utilizes at least 100 formats. He saw HDF5 as the only way to represent these highly complex datasets in a usable, elegant, and consistent way.

➡ A fascinating talk by **Peter Bajcsy** at NCSA concerned a different kind of interoperability: a framework for understanding file format conversions, so that we can make files both backward *and forward* compatible when we have no idea what formats will be used in the future. NCSA has developed a software conversion registry and software that can convert from format A to format X regardless of how many hops (intermediate formats) in the path. They are also researching ways to assess how much information has been lost in the conversion.

The Dresden workshop, which this writer did not attend, had a different set of speakers from Europe, Japan, and New Zealand, and an equally interesting program.

PRISCILLA CAPLAN <pcaplan@ufl.edu> is Assistant Director for Digital Library Services, Florida Center for Library Automation. Among other responsibilities, she directs the Florida Digital Archive, a long-term preservation repository using the DAITSS open source repository applications, and represents FCLA on the Towards Interoperable Preservation Repositories (TIPR) project. Priscilla is the Guest Content Editor for this special preservation-themed issue of ISQ.

Digital Preservation Interoperability Framework workshops
ddp.nist.gov/symposium/home.php

BagIt
wiki.ucop.edu/display/Curation/BagIt

Chronopolis
chronopolis.sdsc.edu/

HDF5
www.hdfgroup.org/HDF5/

MetaArchive Cooperative
www.metaarchive.org/

TIPR: Towards Interoperable Preservation Repositories
wiki.fcla.edu:8000/TIPR

RELEVANT
LINKS