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

CROSSREF. CELEBRATING TEN YEARS OF COLLABORATION

NORTH AMERICA
40 SALEM STREET
LYNNFIELD, MA 01940
UNITED STATES
+1 781 295 0072

EUROPE
CROSSREF UK
3RD FLOOR, 130 HIGH ST
OXFORD, OX1 4DH
UNITED KINGDOM

crossref.org

ISQ

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

NISO
How the information world CONNECTS

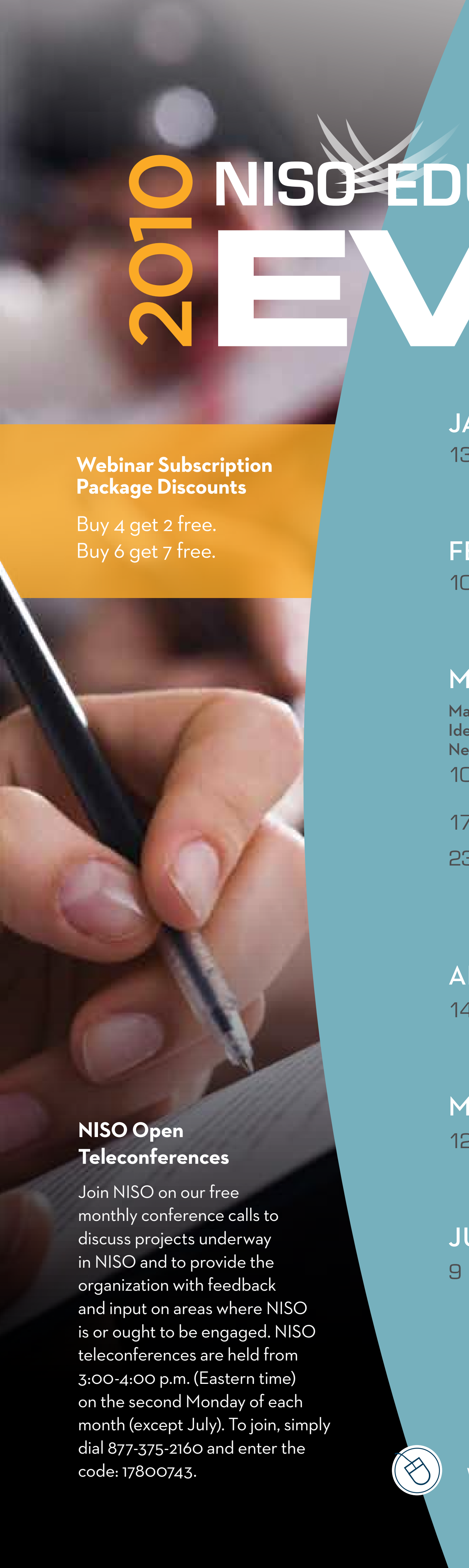
INFORMATION STANDARDS QUARTERLY (ISQ) is a publication by the National Information Standards Organization (NISO). ISQ is NISO's print and electronic magazine for communicating standards-based technology and best practices in library, publishing, and information technology, particularly where these three areas overlap. ISQ reports on the progress of active developments and also on implementations, case studies, and best practices that show potentially replicable efforts.

NISO MANAGING DIRECTOR and ISQ PUBLISHER | Todd Carpenter
ISQ MANAGING EDITOR | Cynthia Hodgson
NISO STANDARDS PROGRAM MANAGER | Karen A. Wetzel
NISO BUSINESS DEVELOPMENT & OPERATIONS MANAGER | Victoria Kinnear
DESIGN | B. Creative Group, Inc.

2010 AD RATES	1 ISSUE	2–3 ISSUES	4 ISSUES
Full page (8.5"x 11")	$375	$350	$325
Half page (4.25" x 5.5")	$250	$225	$200
Back cover (8.5"x 11")	$700	$600	$550
Inside Cover Front (8.5"x 11")	$500	$450	$400
Inside Cover Back (8.5"x 11")	$500	$450	$400

For more information on advertising, visit www.niso.org/publications/isq

CONTENTS



2010 NISO EDUCATIONAL EVENTS

Webinar Subscription Package Discounts

Buy 4 get 2 free.
Buy 6 get 7 free.

JANUARY

13 **Webinar**
From ILS to Repository and Back: Data Interoperability

FEBRUARY

10 **Webinar**
What It Takes To Make It Last: E-Resources Preservation

MARCH

March Two-Part Webinar: Identifiers: New Problems, New Solutions

10 What's in a Name? Latest Developments in Identifiers

17 Content Identification: What's New

23 **In-Person**
Discovery to Delivery: Creating a First-Class User Experience
Atlanta, GA

APRIL

14 **Webinar**
RFID in Libraries: Standards and Expanding Use

MAY

12 **Webinar**
It's in the Mail: Best Practices for Resource Sharing

JUNE

9 **Webinar**
Control Your Vocabulary: Real-World Applications of Semantic Technology

25 **In-Person**
4th Annual NISO/BISG Changing Standards Landscape Forum
Free, open forum at the ALA Annual Conference
Washington, DC

AUGUST

11 **Webinar**
Show Me the Data: Managing Data Sets for Scholarly Content

SEPTEMBER

Two-Part Webinar: Measuring Use, Assessing Success

8 Measure, Assess, Improve, Repeat: Using Library Performance Metrics

15 Count Me In: Measuring Individual Item Usage

OCTOBER

7 **In-Person**
E-Resource Management: From Start to Finish (and Back Again)
Chicago, IL

13 **Webinar**
It's Only as Good as the Metadata: Improving OpenURL and Knowledgebase Quality

NOVEMBER

10 **Webinar**
The Case of the Disappearing Journal: Solving the Title Transfer and Online Display Mystery

DECEMBER

8 **Webinar**
Unprecedented Interaction: Providing Accessibility for the Disabled

NISO Open Teleconferences

Join NISO on our free monthly conference calls to discuss projects underway in NISO and to provide the organization with feedback and input on areas where NISO is or ought to be engaged. NISO teleconferences are held from 3:00-4:00 p.m. (Eastern time) on the second Monday of each month (except July). To join, simply dial 877-375-2160 and enter the code: 17800743.

www.niso.org/news/events

Priscilla Caplan

FROM THE **GUEST CONTENT EDITOR**

In 1997 Terry Kuny, a consultant for the National Library of Canada, prophesied the world was entering a digital Dark Ages. He wrote: "...it is important to know that there are new barbarians at the gate and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever."**

Happily, much that has happened in the following thirteen years makes the prognosis far less grim. We still suffer from an over-abundance of digital information, rapid obsolescence of hardware and software, and increasingly restrictive intellectual property regimes. At the same time, the government, scientific, and cultural heritage sectors have taken the problem to heart and made substantial investments in research and infrastructure to ensure continued access to the human record. A vibrant international community of preservation specialists has moved rapidly from problem definition to the prototyping of solutions, and we observe an increasing emphasis on integrating digital curation and preservation tools into the working environments of libraries, archives, and data centers.

Watching standards evolve in such a young and rapidly growing field has been interesting. Although we often hear it said that premature standardization can have a stultifying effect on experimentation and innovation, the digital preservation community has shown a tremendous thirst for shared specifications of all sorts: frameworks, process models, best practice guidelines, and technical standards. The bible of the preservation domain, *Reference Model for an Open Archival Information System* (OAIS) (ISO 14721:2003) came out of the space science data community but was immediately adopted by cultural heritage institutions. XML file descriptions output by a format identification and validation tool called JHOVE have become de facto standards for format-specific technical metadata, as implementations strive for consistency. The *PREMIS Data Dictionary for Preservation Metadata* had implementations worldwide within two years of being issued, despite having no formal standing as a standard. We see more examples of the drive for common standards in the articles by Angela Dappert and Markus Enders, Andrea Goethals, Carl Fleischhauer and others in this issue of *Information Standards Quarterly*.

We also see how standards development mirrors digital preservation itself in being a highly international, global endeavor. No one country, or even continent, is dominant in leadership. In their *ISQ* article, Robin Dale and Emily Gore reference initiatives led by Canada (InterPARES), the U.K. (DCC Lifecycle Model, DRAMBORA), Germany (Catalogue of Criteria), and the U.S. (TRAC). The articles by Evelyn McLellan and by Kevin DeVorsey and Peter McKinney discuss efforts underway in Canada and New Zealand, respectively. The vast majority of standards efforts have international participation: the group that developed OAIS represents space agencies in 28 nations; the PREMIS Editorial Committee has members from seven countries. Most preservation-related specifications aiming for formal standardization go directly to the International Organization for Standardization (ISO).

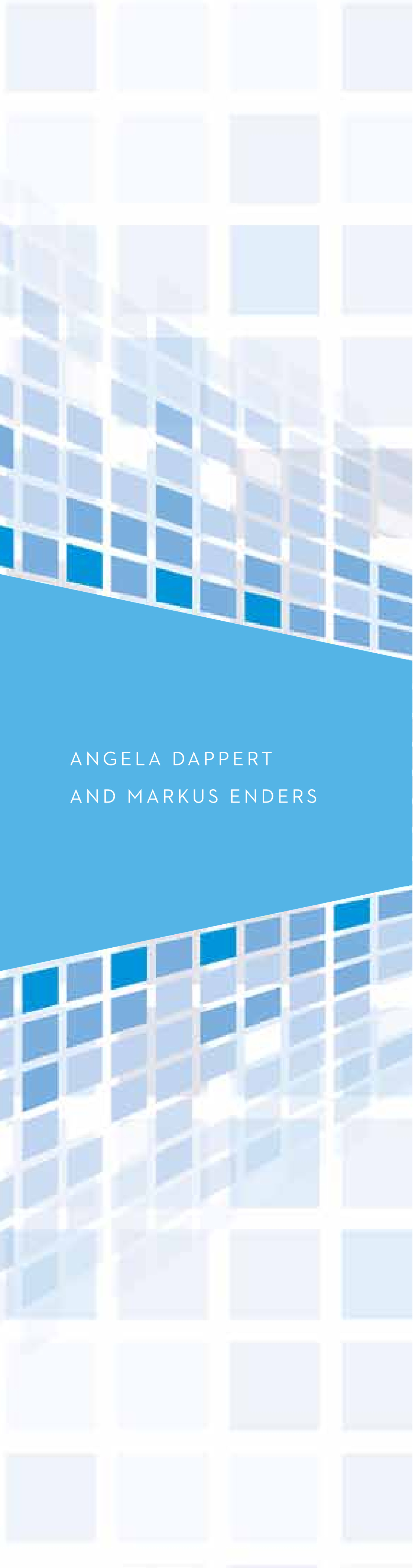
This is not to say, however, that there is global homogeneity of approach and focus. For example, one striking difference between North American and European programs is the amount of attention and effort paid to educating the library and vendor community and involving practitioners at all levels. In the U.S., the National Digital Information Infrastructure and Preservation Project (NDIIPP) has focused on building a community of partners and funding their initiatives. In contrast, the European Planets Project has had broad outreach and training in both theory and practice as a core part of its mission. Digital Preservation Europe (DPE) and the U.K. Digital Curation Centre (DCC) have also had strong core outreach components. The Opinion piece by Mary Molinaro hints that the U.S. situation may improve, which would be a welcome development.

Priscilla Caplan | *Assistant Director for Digital Library Services, Florida Center for Library Automation and ISQ Guest Content Editor*

** Kuny, T. *The Digital Dark Ages? Challenges in the Preservation of Electronic Information*. In: Proceedings of the 63rd IFLA General Conference, Copenhagen, Denmark, August 31- September 5, 1977. Available at: http://archive.ifla.org/IV/ifla63/63kuny1.pdf

DIGITAL PRESERVATION

METADATA STANDARDS



ANGELA DAPPERT
AND MARKUS ENDERS

Valuable scientific and cultural information assets are created, stored, managed, and accessed digitally, but the threat of losing them over the long term is high. Digital media are brittle and short lived. Hardware and software technology continues to evolve at a rapid rate. Changes in organizations and their cultural and financial priorities add risk to continued accessibility and long-term preservation of digital assets. Unlike print-based materials, digital assets cannot survive significant gaps in preservation care.

Digital repositories are computer systems that ingest, store, manage, preserve, and provide access to digital content for the long-term. This requires them to go beyond simple file or bitstream preservation. They must focus on preserving the information and not just the current file-based representation of this information. It is the actual information content of a document, data-set, or sound or video recording that should be preserved, not the Microsoft Word file, the Excel spreadsheet, or the QuickTime movie. The latter represent the information content in a specific file format that will become obsolete in the future.

Preservation policies define how to manage digital assets in a repository to avert the risk of content loss. They specify, amongst other things, data storage requirements, preservation actions, and responsibilities. A preservation policy specifies digital preservation goals to ensure that:

- digital content is within the physical control of the repository;
- digital content can be uniquely and persistently identified and retrieved in the future;
- all information is available so that digital content can be understood by its designated user community;
- significant characteristics of the digital assets are preserved even as data carriers or physical representations change;
- physical media are cared for;
- digital objects remain renderable or executable;
- digital objects remain whole and unimpaired and that it is clear how all the parts relate to each other; and
- digital objects are what they purport to be.

Digital Preservation Metadata

All of these preservation functions depend on the availability of preservation metadata—information that describes the digital content in the repository to ensure its long-term accessibility.

While the Open Archival Information System (OAIS) reference model defines a framework with a common vocabulary and provides a functional and information model for the preservation community, it does not define which specific metadata should be collected or how it should be implemented in order to support preservation goals.

CONTINUED »

The specific metadata needed for long-term preservation falls into four categories based on basic preservation functional groupings:

1 **Descriptive metadata**
Describes the intellectual entity through properties such as author and title, and supports discovery and delivery of digital content. It may also provide an historic context, by, for example, specifying which print-based material was the original source for a digital derivative (source provenance).

2 **Structural metadata**
Captures physical structural relationships, such as which image is embedded within which website, as well as logical structural relationships, such as which page follows which in a digitized book.

3 **Technical metadata for physical files**
Includes technical information that applies to any file type, such as information about the software and hardware on which the digital object can be rendered or executed, or checksums and digital signatures to ensure fixity and authenticity. It also includes content type-specific technical information, such as *image width* for an image or *elapsed time* for an audio file.

4 **Administrative metadata**
Includes provenance information of who has cared for the digital object and what preservation actions have been performed on it, as well as rights and permission information that specifies, for example, access to the digital object, including which preservation actions are permissible.

Even though all four categories are essential for digital preservation, the latter category in particular is often referred to as Preservation Metadata.

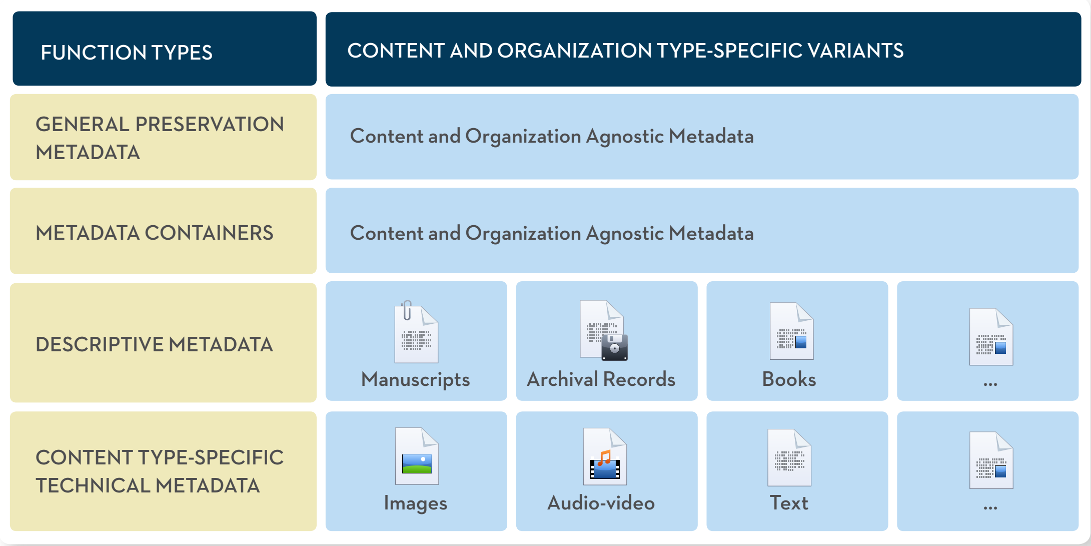
Other analyses and frameworks will use somewhat different categories of preservation metadata. No matter which categories are used, however, they are never clear-cut or unambiguous. A semantic unit can support several preservation functions and, therefore, fall into several categories. For example, the semantic unit *file size* can support both search (e.g., by letting a user search for small images only) and technical repository processes which depend on file size.

The term "semantic unit" is borrowed here from the PREMIS data dictionary. Semantic units are the properties that describe the digital objects and their contexts or relationships between them. The term "metadata element," in contrast, is used to specify how to implement that "semantic unit" in a given metadata implementation specification.

The entities that are described by semantic units are the digital objects themselves, both as abstract, intellectual entities and as physical realizations in the form of renderable or executable file sets. Semantic units can also describe a digital object's hardware, software, and societal environments; rights and permissions attached to them; software and human agents involved in the preservation process; and events that took place during the digital object's life cycle.

Combining Digital Preservation Metadata Specifications

In the early days of digital preservation, there were several uncoordinated efforts to define institution-specific sets of semantic units and metadata elements. These efforts were soon merged into a smaller number



FUNCTION TYPES	CONTENT AND ORGANIZATION TYPE-SPECIFIC VARIANTS			
GENERAL PRESERVATION METADATA	Content and Organization Agnostic Metadata			
METADATA CONTAINERS	Content and Organization Agnostic Metadata			
DESCRIPTIVE METADATA	Manuscripts	Archival Records	Books	...
CONTENT TYPE-SPECIFIC TECHNICAL METADATA	Images	Audio-video	Text	...

FIGURE 1: The Space of Digital Preservation Metadata Efforts

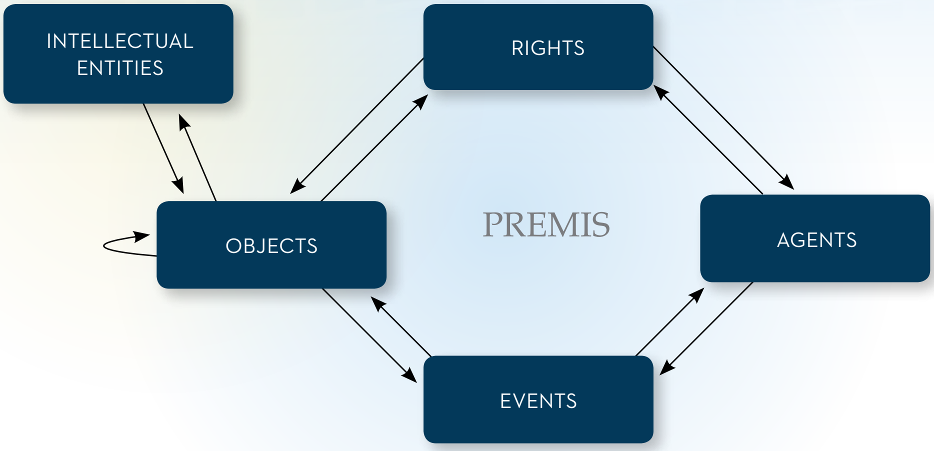


FIGURE 2: The PREMIS Data Model

PREMIS (PREservation Metadata: Implementation Strategies) is one attempt at specifying the semantic units needed to support core preservation functions.

of coordinated international activities that aimed to define sharable preservation metadata specifications. This would ensure interoperability—the ability to exchange amongst institutions and to understand the digital object metadata and its digital content.

A complication was, however, the breadth of metadata needed to support the full range of digital preservation goals. Many years of expertise and effort had already gone into specifying metadata dictionaries or implementation specifications for subsets of the four categories listed above that are also used to support functions outside digital preservation. There was no point in trying to reproduce or outdo this effort. Additionally, it is not possible to define one set of metadata that applies equally to all content types or organization types. Archival records, manuscripts, and library records, for example, require different descriptive metadata; images, text-based documents, and software source code require different technical metadata. Because of this, a number of metadata definition efforts have evolved, both in a content type- or organization type-specific space and a preservation function space. Figure 1 illustrates this in a very simplified way. Several of these initiatives have reached the status of a standard or are de facto standards.

In order to be flexible and apply to a wide range of contexts, general preservation metadata and metadata container specifications try to avoid content and organization specific semantics. For example, general preservation metadata will capture the *file size* of files, since there are no digital representations of content that don't involve at least one file, even if the exact file size may depend on an operating system. It would not, however, capture the *issue number*, which applies to serials but not books, or the *resolution*, which applies to images but not text.

To add specificity, general metadata specifications include extension methods to support content or organization specific metadata. These more specific metadata specifications provide complete sets of semantic units for specific contexts. They provide improved interoperability between independent organizations which share identical contexts; but they may be overly specific and exclude possible other uses. This can stimulate the development of multiple, incompatible metadata solutions to accommodate minor variations in requirements. It is difficult to strike the right balance between generality and specificity. Nonetheless, reusable frameworks with well defined extension points that allow for specific community agreed schemas have been a major advance.

When combining different metadata specifications or when embedding extension metadata, we often find that data models are mismatched or that semantic units overlap. In these cases, it is necessary to decide how to overcome the conflicts. When users make different decisions about how to do this, the interoperability of their metadata suffers. User communities or the bodies that create the metadata specifications can correct for this by specifying best practice guidelines for combining the different metadata specifications. Interoperability can also be improved when users document in metadata profiles how their institution has used a metadata standard for a specific application, including which semantic units and extension schemas have been used for the corresponding items in their data model. If users share their profiles by registering them with a standards editorial board, they may be reused by other potential users with similar content streams, data models, and business use cases.

Descriptive Metadata

Descriptive metadata approaches have been well covered and thoroughly discussed beyond the digital preservation community, and we do not cover them further. This includes both general purpose approaches, such as Dublin Core, and library community approaches, such as MODS and MARC.

CONTINUED »

Semantic Unit 1.5.3: *size*

FIGURE 3: Example PREMIS Semantic Unit

SEMANTIC COMPONENTS	None		
DEFINITION	The size in bytes of the file or bitstream stored in the repository.		
RATIONALE	Size is useful for ensuring the correct number of bytes from storage have been retrieved and that an application has enough room to move or process files. It might also be used when billing for storage.		
DATA CONSTRAINT	Integer		
OBJECT CATEGORY	REPRESENTATION	FILE	BITSTREAM
Applicability	Not Applicable	Applicable	Applicable
Examples	–	2038937	–
Repeatability	–	Not Repeatable	Not Repeatable
Obligation	–	Optional	Optional
CREATION/ MAINTENANCE NOTES	Automatically obtained by the repository.		
USAGE NOTES	Defining this semantic unit as a size in bytes makes it unnecessary to record a unit of measurement. However, for the purpose of data exchange the unit of measurement should be stated or understood by both partners.		

PREMIS makes no assumptions about specific technology, architecture, content type, or preservation strategies. As a result, it is "technically neutral" and supports a wide range of implementation architectures.

Preservation Specific Metadata

Two examples of preservation specific metadata specifications are PREMIS and LMER.

PREMIS (PREservation Metadata: Implementation Strategies) is one attempt at specifying the semantic units needed to support core preservation functions. Core preservation metadata is relevant to a wide range of digital preservation systems and contexts, and it is what "most working preservation repositories are likely to need to know" to preserve digital material over the long term. This includes administrative metadata, but also generic technical metadata that is shared by all content types. It permits the specification of structural relationships if this is relevant for preservation functions, but users may chose to instead use the structural relationships offered by their container metadata specifications, as discussed below.

PREMIS defines a common data model to encourage a shared way of thinking about and for organizing preservation metadata.

The semantic units that describe the entities in this data model (illustrated in Figure 2) are rigorously defined in PREMIS's data dictionary. PREMIS supports specific implementations through guidelines for their management and use and puts an emphasis on enabling automated workflows. It makes, however, no assumptions about specific technology, architecture, content type, or preservation strategies. As a result, it is "technically neutral" and supports a wide range of implementation architectures. For example, metadata could be stored locally or in an external registry (such as a shared file format registry); it could be stored explicitly or known implicitly (e.g., all content in the repository are newspaper articles). PREMIS does not even specify whether a semantic unit has to be implemented through a single field or through more complex data structures. Nonetheless, the PREMIS Editorial Committee maintains an optional XML schema for the convenience of the community.

While PREMIS is very flexible about possible repository-internal implementations, in order to improve interoperability, it is more restrictive on cross-repository information package exchange.

An example PREMIS data dictionary entry for the semantic unit size *is depicted in Figure 3.*

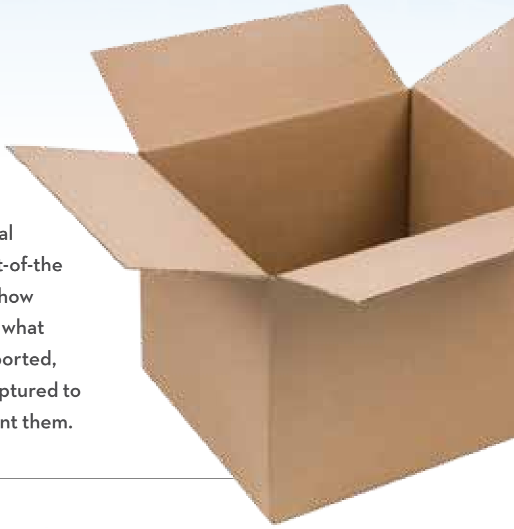
Given the wide range of institutional contexts, PREMIS cannot be an out-of-the box solution. Users have to decide how to model their specific application, what business functions need to be supported, which semantic units need to be captured to support them, and how to implement them. In addition, they need to decide on all metadata that is necessary to manage the content that is not captured in the core preservation metadata.

LMER (Long-term preservation Metadata for Electronic Resources) of the German National Library is an alternative solution to capturing preservation metadata. LMER was designed to meet the requirements of a specific project. Unlike PREMIS it is not a general model for long-term preservation metadata. It implies specific preservation strategies, such as file format migrations, and records detailed information to support this type of preservation action. It enables documenting the provenance of a digital object including tools, reasons, and relationships. As with PREMIS, it includes basic technical metadata, such as checksums and format information. Content type-specific metadata can be embedded using additional schemas such as MIX or TextMD.

LMER's process approach is more workflow oriented than the PREMIS event approach. Any modification to an object is interpreted as a planned process, whereas PREMIS events coincide with the planning that impacts the preserved objects.

Significant Characteristics

When preservation actions are performed on a digital object in its original environment, usually a new digital object is created which is rendered or executed in a new environment. For example, a Word file in its Microsoft rendering environment is migrated to a PDF file in an Adobe rendering environment. With most preservation actions, there is a risk that some characteristics of the original digital object will be lost or modified. In the example migration, one might lose original macros, editing histories, and a degree of interactivity not supported in PDF.



Given the wide range of institutional contexts, PREMIS cannot be an out-of-the box solution. Users have to decide how to model their specific application, what business functions need to be supported, which semantic units need to be captured to support them, and how to implement them.

Significant characteristics reflect business requirements. They capture the characteristics of the original object and environment that need to be preserved by a preservation action. For example, one might wish to specify that for a newspaper collection all pages need to maintain their original margins in a content migration. This requirement guides decisions on which preservation actions should be selected. This specific requirement would, for example, exclude migrations which include cropping within the page edges.

Significant characteristics are a form of preservation metadata that has recently found increased attention. PREMIS supports the capture of simple significant properties for individual digital objects; the PRONOM file format registry project is working on identifying properties that are applicable to file formats; the InSPECT project is working on identifying properties that apply to content types, such as images or e-mails; and the Planets project is investigating advanced significant characteristics and uses them in preservation planning.

Metadata Containers

Digital objects are abstract objects which represent the information entity that should be preserved, accessed, or managed. Metadata containers aggregate their descriptive, administrative, technical, and structural metadata, as well as their physical representations into a single serialization.

Metadata Container Specifications: Since XML is human as well as machine readable, it is the preferred method for specifying metadata containers; it is self-descriptive. The container specifications, however, don't specify a single XML schema containing the complete set of metadata

CONTINUED »

elements. Rather, they are frameworks of high-level elements that define extension points where specific descriptive, administrative, technical, and structural metadata can be embedded. This specific metadata is captured in extension schemas that define the specific metadata elements. It may be physically embedded or reference externally stored metadata.

Structural Metadata: In the analog world, most physical objects are described by a non-hierarchical catalog record. Exceptionally, a catalog may capture the hierarchical containment of parts, such as articles within a serial issue. Digital objects are decomposed to a much finer level of granularity. Even a simple webpage is a complex object. It typically comprises an html file, as well as images, JavaScript, and style sheets. All are required to render the digital object. Additionally, relationships exist between webpages that form a network of objects, allowing users to navigate between them. Each digital object component can be addressed separately—either directly or by following the relationships between components. Their relationships are captured through structural metadata to create one coherent digital object.

Physically, digital objects are represented through files or bytestreams. One digital object may have multiple representations, such as a TIFF and an OCRed text representation of the same newspaper page. Structural metadata relates the abstract object to its physical representations.

Two examples of container specifications are METS and MPEG-21 DIDL.

The **METS (Metadata Encoding and Transmission Standard)** is a specification for exchanging and storing metadata independent of specific project needs.

The only mandatory section in METS is the *structMap* section. Digital objects can be described from different perspectives, resulting in different *structMap* sections. The physical perspective may describe pages, columns, and text areas and their layout relative to each other. The logical perspective may describe sequences, such as the sequence of songs on a CD, or containment, such as the containment of a chapter in a book. These perspectives are captured in separate hierarchical tree structures. Objects in *structMap* sections can be linked to each other. They also can be linked to the file section which describes the corresponding files.

Files in the file section can be organized into one or more file groups. Files may be grouped according to user needs, for example by file format, image resolution, or the intended use of the file (preservation copy, access copy, thumbnail, etc.).

Every object defined in the *structMap* section, as well as every file, may have descriptive or administrative metadata (divided into provenance, source, and technical or rights

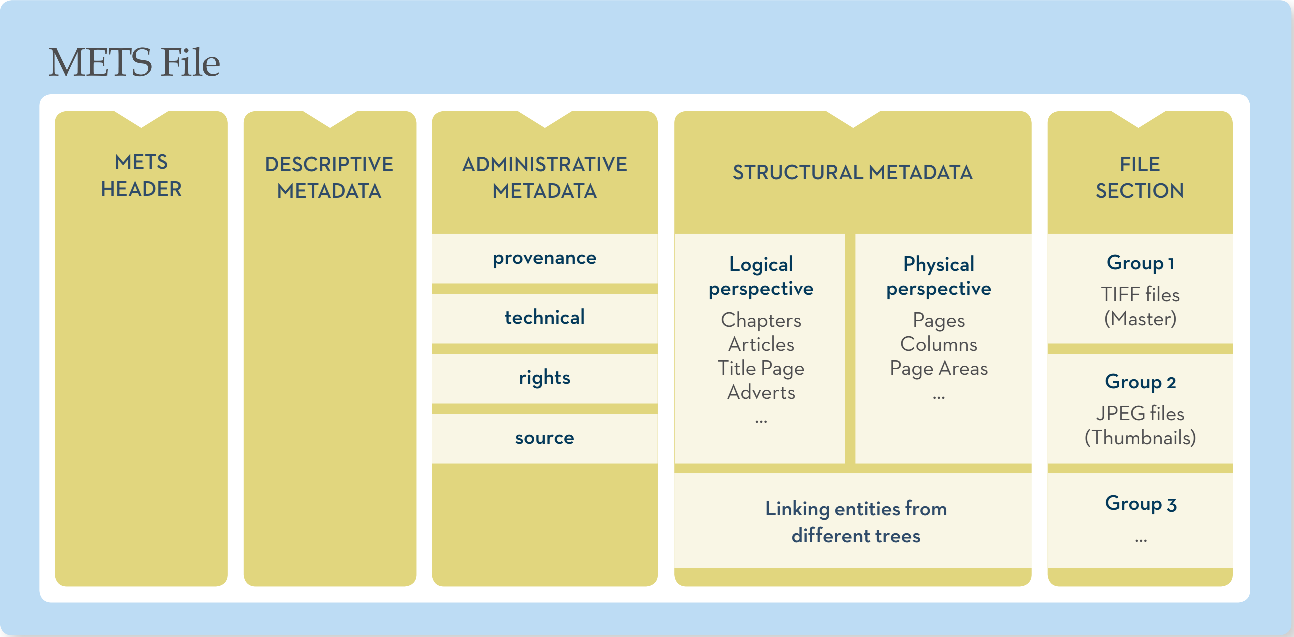


FIGURE 4: The METS Architecture

metadata within METS) describing them outside the *structMap* or file section. Even though METS endorses the use of particular extension schemas, it supports every kind of well-formed XML in these sections. METS uses XML's ID/IDREF linking mechanism for attaching the metadata section to the object. Figure 4 illustrates the METS architecture.

The **MPEG-21** standard has been developed by the Moving Picture Experts Group (MPEG) Committee as an open framework for the delivery and exchange of multimedia objects. It must provide the flexibility required to describe complex audiovisual resources and support any media type and genre. The modular architecture of the MPEG-21 standard allows implementers to pick use case-specific parts of the 12-part standard without losing standard compliance.

Part 2 of this standard is the **Digital Item Declaration Language (DIDL)**. DIDL uses five basic concepts for describing complex digital objects. The semantics of these concepts are more abstract than the sections in METS. *Containers* can group containers and/or items. An *item* can group further items or components. A *component* groups resources. All *resources* within the same component are regarded as semantically equivalent. DIDL defines a resource as an individual bytestream that contains the actual content of an item and can either be embedded into the DIDL description or referenced.

DIDL only defines the structure of a complex object. Any additional descriptive or administrative metadata about a container, item, or component must be stored in a metadata wrapper, called a *descriptor*. The MPEG-21 Rights Expression Language (REL) in Part 5 and the Digital Item Identification Language (DII) in Part 3 of the standard can be used to capture some of this metadata. Additionally, a descriptor may contain any non-MPEG-21 XML structure to capture preservation metadata.

MPEG-21 DIDL defines a conceptual data model and its representation as an XML bytestream. The container, item, component, resource, and descriptor objects are represented as nested XML elements. Therefore, an ID/IDREF linking mechanism for linking different sections is, unlike in METS, not necessary. Unlike METS, DIDL provides few attributes for capturing technical or descriptive metadata. Figure 5 illustrates the MPEG-21 DIDL architecture.

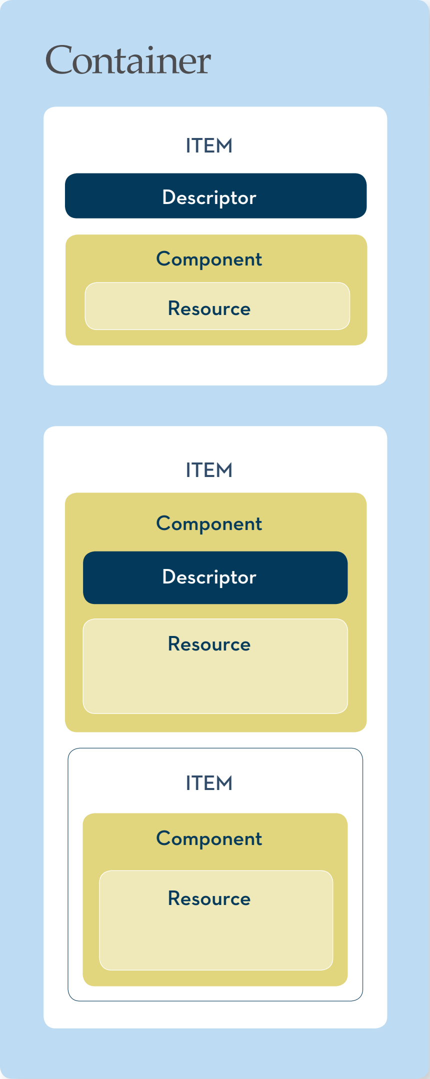


FIGURE 5: The MPEG-21 DIDL Architecture

Content Type-Specific Technical Metadata

Technical metadata may be specific to a content type, such as raster or vector image, sound, video, text, spreadsheet, or e-mail.

Some content type-specific metadata is essential for rendering a digital object representation. For example, it is essential to know the sample rate of digital audio data, or the width, height, and color depth of an image.

Some file formats enable the capture of technical, and other, metadata within their files, which has the advantage of keeping the files self-descriptive. However, by extracting and storing metadata explicitly we may also benefit. Separate metadata can:

- be kept small and processed efficiently;
- be distributed separately;
- have different access rights and licensing arrangements than the content;
- help to account for the whole life cycle of digital objects;
- have its description standardized across file formats; and
- be managed and preserved by preservation systems.

> Some file formats enable the capture of technical, and other, metadata within their files, which has the advantage of keeping the files self-descriptive. However, by extracting and storing metadata explicitly we may also benefit.

CONTINUED »

Content type-specific technical metadata is typically introduced through an extension schema within container formats such as METS or MPEG21 DIDL.

Two examples of content type-specific metadata are the ANSI/NISO Z39.87 standard and the textMD specification.

The **ANSI/NISO Z39.87** standard, *Data Dictionary –Technical Metadata for Digital Still Images*, defines semantic units to describe digital raster images. The standard does not prescribe a serialization. But, in partnership with NISO, the Library of Congress maintains an XML Schema called **MIX (Metadata for Images in XML Schema)** that is widely used by content creators and in the digital preservation community. Tools, such as JHOVE, are available to extract technical metadata from image files and export the metadata as MIX serialization.

Like the Z39.87 standard, MIX defines four sections of metadata:

1 **Basic Digital Object Information:** Basic non-content type-specific metadata such as file size, checksums, and format information.

2 **Basic Image Information:** Metadata that is required to render an image, including the compression algorithm and the image dimensions.

3 **Image Capture Metadata:** Metadata about the image capturing process, such as the scanning device, settings, and software used in the process.

4 **Image Assessment Metadata:** Metadata important for maintaining the image quality. Information in this section is necessary to assess the accuracy of output. This includes color information (such as white points and color maps) and resolution information.

TextMD is a technical metadata specification for text-based digital objects expressed as an XML schema. The schema provides elements for storing the encoding and character information such as *byte order*, *linebreaks*, *character set*, and information about the technical environment in which the text was created.

It may also store information about the technical requirements for printing or rendering the text on screen. This includes information about sequences and page ordering and may therefore overlap with information stored as structural metadata in the metadata container. While textMD is attached to text files, individual document pages may additionally be defined as distinct objects with their own metadata.

Metadata Exchange

Preserving digital content is a collaborative effort. Organizations which are running a preservation repository may want to share content with selected partners to provide distributed preservation solutions. These preservation solutions must exchange complex objects between heterogeneous preservation systems.

Preserving digital content is a *collaborative* effort. Organizations which are running a preservation repository may want to share content with selected partners to provide distributed preservation solutions. These preservation solutions must exchange complex objects between heterogeneous preservation systems.

The **TIPR (Towards Interoperable Preservation Repositories)** project develops a prototype for distributing content between three different partners who are running technically heterogeneous repository systems with distinct data models. The common transfer format for the information package is based on METS and PREMIS as defined in the TIPR Repository Exchange Package (RXP). In order to handle the different data models manifested in the complex objects from other partners, each repository must understand the other repository's data model. The de facto standards METS and PREMIS proved to be flexible enough for transmitting the information packages between repositories.

Conclusion

This article introduced metadata for digital preservation and argued why it is needed. It outlined the space of different metadata specifications and alluded to the problems inherent in defining and combining a small, but comprehensive set of standards.

Currently, few metadata specifications contributing to digital assets' long-term preservation are sanctioned by national or international standards bodies. Some, like PREMIS or METS, have the status of de facto standards with well-defined community processes for maintaining and updating them. While communities have a strong desire for long-lasting, stable metadata standards, they continue to evolve as the number of repository implementations and applications grows. Experience remains too limited to set a preservation metadata standard in stone.

In addition to strong growth in practical experience, research and technology development projects, such as the EU co-funded Planets project, have added substantially to our fundamental understanding of the preservation metadata space. They have brought us closer to end-to-end digital preservation solutions that test the flow of preservation metadata across multiple digital preservation services. This combination of practical experience and renewed fundamental exploration contributes to a growing understanding of digital preservation metadata.

| FE | doi: 10.3789/isqv22n2.2010.01

ANGELA DAPPERT <Angela.Dappert@bl.uk> is Digital Preservation Manager and **MARKUS ENDERS** <Markus.Enders@bl.uk> is Technical Architect, both at The British Library <www.bl.uk>, where they, amongst other tasks, collaboratively develop metadata profiles for the digital library system. Both serve on the PREMIS Editorial Committee, and Markus also serves on the METS Editorial Board.

RELEVANT LINKS

ANSI/NISO Z39.87
www.niso.org/standards/z39-87-2006/

InSPECT
www.significantproperties.org.uk/

LMER
www.d-nb.de/eng/standards/lmer/lmer.htm

METS
www.loc.gov/standards/mets/

MIX
www.loc.gov/standards/mix/

MPEG-21 DIDL
www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=35366

OAIS
www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf

Planets
www.planets-project.eu/

PREMIS
www.loc.gov/standards/premis/

PRONOM
www.nationalarchives.gov.uk/PRONOM/Default.aspx

textMD
www.loc.gov/standards/textMD/

TIPR Repository Exchange Package
wiki.fcla.edu:8000/TIPR

ROBIN L. DALE AND EMILY B. GORE

Process Models and the Development of TRUSTWORTHY DIGITAL REPOSITORIES

In the past 15 years, there has been much effort to address the long-term preservation of digital assets, including the establishment of standards, related guidance, and best practices. In this article, the authors will give an overview of process models for preservation, including OAIS, InterPARES, and the DCC Curation Lifecycle Model, and the relationship of those process models to the development of standards related to trustworthy repositories. A discussion of work towards developing standards and best practices to establish trustworthy repositories begins with the seminal documents *Preserving Digital Information* and *Trusted Digital Repositories* (TDR) continues through currently used de facto standards TRAC, DRAMBORA, and nestor, and concludes with certification-related standards emerging from the OAIS family of standards. Process models and their intersections with efforts to provide guidance and set standards for trustworthy repositories guide the work of practitioners charged with long-term digital asset management across many disciplines.

Process Models

Process modeling is the activity of representing processes of a community, often so that current processes may be understood, analyzed, and improved. Process models are typically descriptive, prescriptive, and explanatory. The development of process models often begins by looking at the way processes have historically been performed and improvements for efficiency and effectiveness were determined. Process models then establish rules and guidelines that lead to desired process performance and provide explanations about the rationale of processes.

Early discourse about digital preservation tended to focus on specific technological strategies for digital files, but left important issues unaddressed. In developing process models for digital preservation, the community was forced to model and document the entire context in which those digital files existed, revealing overarching requirements for the infrastructure, supporting information models, processes, and systems in which they exist.

Open Archival Information System (OAIS)

In the early 1990s, the Consultative Committee for Space Data Systems (CCSDS) initiated work aimed at developing formal standards for the long-term storage of digital data generated from space missions. As described by Lavoie, this work was initially hindered because in early research, the CCSDS found no widely-accepted framework that could serve as a foundation for standards-building activities: nothing that established shared concepts and terminology associated with digital preservation, characterized the basic functions constituting a digital archiving system, or defined the important attributes of the digital information objects towards which preservation efforts could be directed. In 1995, the CCSDS began development of a framework that would serve the broadest constituency possible, incorporating relevant work from communities outside of the space data community including the seminal work, *Preserving Digital Information*, from the Task Force on Archiving of Digital Information.

Since the release of the CCSDS' draft OAIS reference model in 1999, archival repository systems worldwide have used OAIS as a benchmark and as the chief process model for the preservation of digital assets. The reference model

provides a common conceptual framework describing the environment, functional components, and information objects within a system responsible for the long-term preservation of digital materials. OAIS as a process model does not prescribe standards or technical architectures for archives or repositories; rather it gives a framework for further, more granular standards development and establishes an ontology for communication among repositories.

In 2003, the *Reference Model for an Open Archival Information System* was formalized and published as ISO 14721, paving the way for the development of future digital preservation standards work. The OAIS included a Roadmap for follow-on standards which led to the development of related process models. Follow-on or related standards development emerged including the *Producer-Archive Interface Methodology Abstract Standard* (PAIS) and the *PREMIS Data Dictionary for Preservation Metadata*. An additional standard was planned for the "accreditation of archives" but because of ongoing, parallel work, it was agreed that RLG and National Archives and Records Administration (NARA) would take this particular topic forward.

InterPARES

While OAIS was being developed, a process model for the long-term preservation of electronic records, InterPARES, was also in development. InterPARES, the International Research on Permanent Authentic Records in Electronic Systems, focuses on a model for ensuring the preservation, accuracy, reliability, and authenticity of electronic records. In Phase 1 (1999-2001), InterPARES work included the development of activity models for the selection and preservation functions, and created a framework for requirements for assessing and maintaining authenticity of electronic records. Benchmark requirements supporting the presumption of authenticity as well as baseline requirements supporting the production of authentic copies of electronic records were also developed during this phase and were documented in the InterPARES *Preserve Electronic Records* model. While ensuring compliance with the OAIS model, the *Preserve Electronic Records* model defines processes specifically related to the preservation and delivery of authentic electronic records, and focuses only on essential preservation-related tasks. In Phase 2 (2002-2007), InterPARES shifted focus to newer kinds of electronic records: those which are dynamic, interactive, and experiential. The goal was to develop understanding surrounding their creation, maintenance, and preservation. Additional developments in this phase included methods for creating, maintaining, and preserving accurate, authentic, and reliable records in the arts, sciences, and government. Phase 3 (2007-2012) is currently underway and focuses on the movement of theory into practice through constituent adoption and education.

Digital Curation Centre Curation Lifecycle Model

A more recent model, the DCC (Digital Curation Centre) Curation Lifecycle Model, provides a graphical overview for the successful curation and preservation of digital assets from concept or receipt. The model aims to illustrate the steps or high-level processes necessary for long-term preservation, and is designed to be used in conjunction with relevant standards to plan curation and preservation activities to different levels of granularity. The DCC asserts that the

CONTINUED »

lifecycle model is intended to complement other models, like OAIS and InterPARES. Because of its intentional high-level overview, "workflow design, management issues, identification of processes and use of best practice can all be enhanced through the application of standards such as OAIS." The model defines three levels of preservation actions: full lifecycle, sequential, and occasional and points to the adherence of established best practices and standards for all levels of action. The DCC encourages use of the model as a training tool for data creators, data curators, and data users; to organize and plan resources; and to help organizations identify risks to their digital assets and plan management strategies for their successful curation.

A "trusted digital repository" is one whose mission is to provide reliable, long-term access to managed digital resources for its designated community, now and in the future.

From Process Models to Certified Digital Repositories

Having set the stage for the development of digital preservation frameworks and process models, *Preserving Digital Information* arguably also sets the stage for "trustworthy repositories" in its seminal work. "A critical component of the digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections…" In its recommendations, the Task Force articulated a need for "a process for certification of digital archives…to create an overall climate of trust about the prospects of preserving digital information." At the time, two potential models were recognized: an audit model based on those used to certify official depositories of government documents and a standards model where "participants claim to adhere to standards that an appropriate agency has certified as valid and appropriate; consumers then certify by their use whether the products and services actually adhere to the standards." Yet formal standards and well-accepted practices for digital preservation were slow to develop in the five years following the publication of *Preserving Digital Information*. Those that did emerge tended to be opposite ends of the standards spectrum: high-level process models and frameworks (OAIS, InterPARES) or more granular standards that addressed core parts of the digital preservation process (PAIS, PREMIS, etc.). The process models lacked the granularity required for an auditable certification process; individual, emerging standards lacked a framework for what constituted a trustworthy repository; and the community remained unable to come to a collective agreement on an exact definition of "trusted archives" as called for by the task force.

Defining Trustworthy Digital Repositories (TDRs)

In March 2000, RLG and OCLC began work to establish attributes of a digital repository for research organizations, building on and incorporating the then-emerging OAIS reference model. Representatives from libraries, archives, and data archives were charged to reach consensus on the characteristics and responsibilities of trusted digital repositories for large-scale, heterogeneous collections held by cultural organizations. The resulting work, *Trusted Digital Repositories: Attributes and Responsibilities*, articulated a framework of attributes and responsibilities for trusted, reliable, sustainable digital repositories capable of handling the range of materials held by large and small research institutions. It also defined a "trusted digital repository" as one whose mission is to provide reliable, long-term access to managed digital resources for its designated community, now and in the future. Inherent in this definition is the concept that preservation and access are inextricably linked but the framework was broad enough to accommodate different situations, architectures, and institutional responsibilities.

Jantz and Giarlo noted that a particular value of the TDR report was the concept that a "trusted digital repository" was based on two major requirements: "1) the repository with associated policies, standards, and technology infrastructure will provide the framework for doing digital preservation, and 2) the repository is a trusted system, i.e., a system of software and hardware that can be relied upon to follow certain rules." The

This concept is based on two major requirements:

1. the repository with associated policies, standards, and technology infrastructure will provide the framework for doing digital preservation, and
2. the repository is a trusted system, i.e., a system of software and hardware that can be relied upon to follow certain rules.

trusted system concept—that long-term digital preservation could not occur in a vacuum but instead existed within a larger organizational ecosystem that played key roles, as well as represented key vulnerabilities in the process—was an important step towards identifying trustworthy aspects of digital repositories. The document proved useful for institutions grappling with the long-term preservation of cultural heritage resources and was used in combination with the OAIS as a digital preservation planning tool. As a framework however, the TDR report concentrated on high-level organizational and technical attributes and only discussed potential models for digital repository certification. It refrained from being prescriptive about the specific nature of rapidly emerging digital repositories and archives and instead reiterated the call for certification of digital repositories, recommending the development of a certification program and the articulation of auditable criteria.

Developing Metrics for Certification

In 2003, RLG and the National Archives and Records Administration (NARA) created a joint task force to specifically address digital repository certification. The goal was to produce certification criteria and delineate a process for certification applicable to a range of digital repositories and archives. The membership of the RLG-NARA Task Force on Digital Repository Certification reflected that diversity, with practitioner-members from each of those organization types. All were chosen because of their experience in building and managing digital repositories. Continuity with earlier efforts was ensured by including members who had played active roles in the development of the OAIS standard and TDR report.

Beginning from a base of practitioner experience and leveraging concepts from existing documentation and standards for related types of certification (the ISO 9000 family of standards relating to organization and system management; ISO 17799 for data security and information management systems; the US Department of Defense Standard DoD 5015.2 (2002) for Records Management Applications, and many others), criteria were established and vetted using an iterative process. After two years, an audit tool comprising 88 metrics had been shaped and was released in draft as *An Audit Checklist for the Certification of Trusted Digital Repositories*.

A valuable public comment period brought important suggestions for improvement to the *Audit Checklist*, including the call for not only characteristics of a trusted digital repository, but also ways in which the presence of the attributes can be demonstrated and their qualities measured (see Ross and McHugh). By its publication, potential complexities of a formal audit and certification process were highlighted and questions were raised about applicability for existing "digital archives" of content. At a time when most digital repositories were in the developmental stage, there was arguably an equal if not greater need for a planning/development tool for trusted repositories than a need for formalized audit and certification of digital repositories. How could a "best practice" audit tool be used to encourage and direct repository development without overwhelming institutions with nascent repositories? Was the *Audit Checklist* necessary and relevant for all digital repositories? Should regional needs or laws drive the development of several checklist versions? Could the checklist be easily used for self-assessment? And how would or could the *Audit Checklist* be applicable to repositories and digital content services that were established long before the *Audit Checklist* was developed?

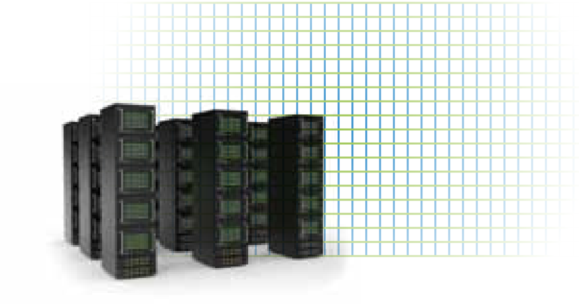
The task force and other organizations considered those questions. The result was not only significant redevelopment of the *Audit Checklist*, but the development of two additional audit and criteria tools by two other organizations. The final phase of certification standards development saw an increase in interest, organizational sponsorship, and organizational participation. Two additional certification activities went into development in 2004–2005 and led to the release of complementary metrics for trustworthy repositories. Additionally, the three groups worked together to produce principles for minimum requirements for trustworthy digital preservation repositories.

1 The nestor Catalogue of Criteria for Trusted Digital Repositories

In December 2004, the German *nestor* project (Network of Expertise in Long-term STOrage of Digital Resources) set up the *nestor* Working Group on Trusted Digital Repository Certification to define a first catalog of criteria for trustworthiness and to prepare for the certification of digital repositories in accordance with nationally and internationally coordinated procedures. The aim of the project was to "establish a net of trustworthiness" in which long-term digital archives can function in various environments by formulating criteria that could be used for a broad spectrum of digital long-term repositories. Similar to the goals of the RLG-NARA task force, there was also a desire to provide information and self assessment assistance with the design, planning, and implementation of digital repositories.

Beginning with a small-scale survey on recent standards and usage within digital repositories within German institutions, the working group followed up with a public workshop in June 2005 and an expert round table in March 2006. Version 1.0 of the *Catalogue of Criteria for Trusted Digital Repositories* was released in June 2006. Comprising abstract criteria, enhanced with examples and explanations, the *Catalogue of Criteria* encompassed international standards but focused on applications in Germany. The central concepts

CONTINUED »



The aim of nestor was to "establish a net of trustworthiness" in which long-term digital archives can function in various environments by formulating criteria that could be used for a broad spectrum of digital long-term repositories.

driving the criteria include trustworthiness, as well as the concept that implementation of any certification process is a multi-step process for repositories and must be iterative. The application principles developed by *nestor*—Documentation, Transparency, Adequacy, and Measurability—were later adapted along with the Digital Curation Centre's needs for evidence in the RLG-NARA task force work. Today, the *nestor Catalogue of Criteria for Trusted Digital Repositories* continues to be in use in Germany in concert with training tools developed by the working group.

2 DRAMBORA: Digital Repository Audit Method Based on Risk Assessment

Developed jointly by the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) is intended to facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses, and recognize their strengths. Borne out of a DCC repository assessment project, the initial basis for assessment was rooted in TRAC audit metrics (see #3) but was designed specifically with self-assessment in mind. DRAMBORA is a methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities, and assets before identifying, assessing, and managing the risks implicit within their organization. This method and the accompanying tool focus on organizations willing to perform a self-assessment to get an overview of the risks in their organization.

DRAMBORA focuses on risk management and asserts that the role of the curator or repository manager is to manage risks. Now available as an online, interactive toolkit, DRAMBORA defines six stages within the risk management process. Through the process of self-assessment, repository managers become aware of shortcomings and greatest risks. A systematic process guides the auditor to identify risks to long-term preservation of repository content, and then scores each risk as a product between the likelihood of the risk occurring with the impact associated with that event. Mitigation of the risks can then be prioritized in descending order of the score so that risks can be effectively managed.

3 TRAC: Trusted Repositories Audit & Certification: Criteria & Checklist

During the final phase of metrics development, the RLG-NARA Task Force on Digital Repository Certification was fortunate to obtain valuable alliances with the then-new Digital Curation Centre, as well as colleagues in Germany directing the *nestor* project. A critical alliance with the Center for Research Libraries (CRL) also emerged. In 2005, the Center for Research Libraries was awarded a grant by The Andrew W. Mellon Foundation to develop the procedures and activities required to audit and certify digital archives. The CRL Certification of Digital Archives Project worked closely together with the RLG-NARA task force to redevelop the audit metrics and provided critical opportunities to develop and test the audit process itself. This practical testing, along with the DCC test audits that led to the development of DRAMBORA, contributed greatly to filling the gaps identified in the earlier draft, *Audit Checklist for the Certification of Trusted Digital Repositories*.

The final version of *TRAC* was published in February 2007 with 84 criteria broken out into three main sections: Organizational infrastructure; Digital object management; and Technologies, technical infrastructure, and security. It provides tools for the audit, assessment, and potential certification of digital repositories; establishes the documentation requirements for audit; delineates a process for certification; and establishes appropriate methodologies for determining the soundness and sustainability of digital repositories.

It currently serves as a de facto standard for repository audit and is being actively used by organizations as both a planning and self assessment tool. Additionally, it continues to serve as the basis of further CRL audit and certification work, including the National Science Foundation-funded project, Long-Lived Digital Collections. Currently, two repositories of interest, Portico and HathiTrust, have agreed to undergo CRL audits. Based on its recent audit findings, CRL has certified Portico as a trustworthy digital repository for the CRL community.

Ongoing Standards Development for Trustworthy Digital Repositories

After the publication of TRAC, the CCSDS working group responsible for OAIS-related standards (now called Mission Operations and Information Management Services or MOIMS) shepherded the TRAC certification metrics back into the CCSDS/ISO standards process. The MOIMS Repository

Audit and Certification (MOIMS-RAC) Working Group has endeavored over the last three years to formalize repository audit and certification metrics and continue the growth of the OAIS family of standards as envisioned at the outset of the OAIS work. Currently, two major contributions are in the standards process:

- *Audit and Certification of Trustworthy Digital Repositories* (currently designated CCSDS 652.0-R1, October 2009) is a draft standard that articulates the audit and certification criteria for trustworthy digital repositories. It is in the balloting and revision process and expected to be released very soon as the new international standard for certification.

- *Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories* (Draft Recommended Practice CCSDS 000.0-R-0, Red Book) is meant primarily for those setting up and managing the organization performing the auditing and certification of digital repositories. Currently, ISO/IEC 17021, *Conformity Assessment Requirements for Bodies Providing Audit and Certification of Management Systems*, is the international standard that prescribes criteria for audit and certification agencies' work. The new CCSDS standard will incorporate new requirements and guidance for agencies to be accredited as complying with ISO/IEC 17021 with the objective of auditing and certifying candidate Trusted Digital Repositories (TDR).

With the formalization of these two documents, the standardization process for trustworthy digital repositories will have completed its first cycle. Not unlike the DCC's Curation Lifecycle Model, this cycle of understanding and standardization will continue as an iterative process. With a stable base of a process model, relevant standards and best practices for individual parts of the process will continue to be developed as the community's experience with and expertise in digital preservation grows. | FE | doi: 10.3789/isqv22n2.2010.02

ROBIN L. DALE <robin.dale@lyrasis.org> is the Director of Digital and Preservation Services for Lyrasis and as former RLG Program Officer was the principal author of *Trusted Digital Repositories: Attributes and Responsibilities, An Audit Checklist for the Certification of Trusted Digital Repositories*, and TRAC. Dale also served as a member of the OAIS, PREMIS, and ANSI/NISO Z39.87, *Technical Metadata for Digital Still Images*, working groups.

EMILY B. GORE <egore@clemson.edu> is the Head of Digital Initiatives & Information Technology for the Clemson University Libraries. Gore is the current co-chair of the ALA ALCTS-PARS Digital Preservation Interest Group, a sustaining member of the MetaArchive Cooperative, and an active repository manager.

RELEVANT LINKS

An Audit Checklist for the Certification of Digital Repositories. RLG and NARA, August 2005.
www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file2416.pdf

Catalogue of Criteria for Trusted Digital Repositories. nestor Working Group on Trusted Repositories Certification, December 2006.
files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf

The DCC Curation Lifecycle Model
www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

DRAMBORA
www.repositoryaudit.eu/

InterPARES Project
www.interpares.org/

Jantz, Ronald and Michael J. Giarlo. Digital Preservation Architecture and Technology for Trusted Digital Repositories. D-Lib Magazine, 11(6), June 2005.
www.dlib.org/dlib/june05/jantz/06jantz.html

Lavoie, Brian. The Open Archival Information System Reference Model: Introductory Guide. DPC Technology Watch Report 04-01, 2004.
www.dpconline.org/vendor-reports/download-document/91-introduction-to-oais.html

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

Producer-Archive Interface Methodology Abstract Standard
public.ccsds.org/publications/archive/651x0b1.pdf

Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data System, January 2002.
public.ccsds.org/publications/archive/650x0b1.pdf

Rolland, Colette and C. Thanos Pernici. A Comprehensive View of Process Engineering. In: Proceedings of the 10th International Conference CAiSE'98. B. Lecture Notes in Computer Science 1413, 1998.
ftp://sunsite.informatik.rwth-aachen.de/pub/CREWS/CREWS-98-18.pdf

Ross, Seamus and Andrew McHugh. The Role of Evidence in Establishing Trust in Repositories. D-Lib Magazine, 12(7/8), July/August 2006.
www.dlib.org/dlib/july06/ross/07ross.html

Ross, Seamus and Andrew McHugh. Preservation Pressure Points: Evaluating Diverse Evidence for Risk Management.
www.repositoryaudit.eu/images/PreservationPressurePoints.pdf

Task Force on Archiving of Digital Information. Preserving Digital Information. Council on Library and Information Resources, pub 63, May 1, 1996.
www.clir.org/pubs/reports/pub63watersgarrett.pdf

Trusted Digital Repositories: Attributes and Responsibilities. RLG-OCLC, 2002.
www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf

Trustworthy Repositories Audit & Certification: Criteria and Checklist. OCLC and CRL, version 1.0, February 2007.
www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

LIZ BISHOFF

DIGITAL PRESERVATION PLAN

Ensuring Long Term Access and Authenticity of Digital Collections

For nearly two decades libraries and cultural heritage organizations have been fulfilling our role as stewards of digital resources by acquiring and reformatting analog collections into digital format and by making the digital resources available to our respective communities to meet their information and education needs.

During that time, millions of digital resources have been created, however little or no thought has been given to the long-term access to these resources. Yet when professionals across the cultural heritage community are surveyed, 78.4% respond that they expect to provide access to these collections for more than 10 years, 2.7% planned to provide access for less than 10 years, while 18.9% didn't know how long they would provide access. (Participants in 2006-2007 NEDCC sponsored, NEH funded Stewardship of Digital Asset (SODA) workshops completed a pre-workshop survey. This survey asked the participants a variety of questions regarding their digital programs. This author was one of the faculty members who have been cumulating data from the 110 institutions who participated in the SODA surveys.) At the same time only 20.7% reported that they had a digital preservation plan. Among the same group, 48% indicated that they planned to become a Trusted Digital Repository.

While the current economy may delay implementation of digital preservation programs, development of digital preservation plans can begin at anytime, allowing the organization to develop the foundation and knowledge required to develop a funding proposal for the digital preservation program. A digital preservation plan is the organization's public statement regarding its commitment to preserve its digital collections through the development and evolution of a comprehensive digital preservation program. The plan will provide the mission, specific goals and objectives, and policies and procedures. It will define the preservation strategies, standards, digital content depositors, staffing, funding, roles and responsibilities, and the users. The digital preservation plan is based on two key documents: *Trusted Digital Repositories: Attributes and Responsibilities* (2003) and the *Open Archival Information System (OAIS) Reference Model* (ISO 14721:2003).

Components and objectives

Digital preservation plans should include the following components:

1. Rationale for digital preservation
2. Statement of organizational commitment
3. Statement of financial commitment
4. Preservation of authentic resources and quality control
5. Metadata creation
6. Roles and responsibilities
7. Training and education
8. Monitoring and review

CONTINUED »

The plan should be a collaborative effort of the Digital Preservation team. The team may involve members of your digital library team including your digital librarian/digital archivist, collection curator/s, preservation librarian, metadata librarian, and IT manager. Additional participants can include legal counsel, financial manager, a representative from senior management, and other appropriate stakeholders.

The plan should support the following objectives:

- Ensure the preservation of and continued access to born digital and digitally reformatted materials.
- Ensure the preserved materials are authentic.
- Preserve physical media from damage and deterioration through appropriate environmental controls.
- Reverse damage, where possible.
- Change format of digital materials to preserve their intellectual content if necessary.

1 Rationale

Why are you creating a digital preservation program? The digital preservation plan should include the rationale for the program. The statement can be simple and straight forward, similar to the one developed by Yale University Library: "Yale University Library Digital Preservation Policy supports the preservation for digital resources that are within the Library's collections." Alternatively, the plan can incorporate statements reflecting the goals of digital preservation; i.e., "…establish centralized responsibility for ensuring continuing access to digital image collections over time…centralized responsibility will facilitate the long term use of digital resources in the most cost effective manner."

2 Organizational Commitment

Organizational commitment is demonstrated through incorporation of statements of support of digital preservation within the organization's mission or mandate. The SODA survey found that 31% of the institutions had specific statements within their mission, while 53.1% did not. Columbia University Library's plan states that "digital resources are part of the CUL collections and subject to the same criteria for selection and retention and decisions as other media…" Other plans provide more specific explanation of their organizational commitment, for example the University of Michigan's Inter-University Consortium for Political and Social Research (ICPSR) includes in their plan both a mandate statement and objectives. Their mandate includes three components: scholarly commitment, membership services, and contractual obligations and grants.

Based on the SODA survey, cultural heritage organizations have placed emphasis on developing policies with equal emphasis on rights management and mission, and emergency preparedness and collection development, while digital preservation is lagging significantly as shown in Figure 1.

POLICY	YES	NO
MISSION	31%	53.1%
COLLECTION DEVELOPMENT	28%	38.8%
EMERGENCY PREPAREDNESS	27%	44.1%
PRESERVATION	21.6%	40.5%
RIGHTS	32.4%	40.5%

Figure 1: Policies in place in cultural organizations

The organization should include a succession plan in the preservation plan. The plan must identify for their users the program's strategy in the event that the digital preservation program is no longer able to support its preservation commitments.

3 Financial Commitment

Financial commitment may be one of the more challenging areas for organizations to address, particularly for those associated with government entities that operate on an annual appropriation. The promise a trusted steward makes cannot be undertaken with one year increments. Technology planning, staff training, software development or acquisition, and legal agreements all are multi-year commitments. However, the financial planning structure rarely supports multi-year planning. Nonetheless, the digital preservation plan must make every effort to consider how to address the financial sustainability of the digital preservation program. Components of the financial commitment may include institutional commitment, legislative mandate (if there is financial support), and membership structure (if the digital preservation program is a collaborative or based on a subscription program),

Financial commitment may be one of the more challenging areas for organizations to address, particularly for those associated with government entities that operate on an annual appropriation.

fund raising and grant programs, and fees and other revenue sources.

Lastly, financial sustainability should require collaborative initiative. These may include collaboration with other digital repositories, data producers, digital preservation programs, standard-setting bodies, and commercial organizations working in the area of digital programs.

4 Preservation Strategies

Many preservation plans will provide a brief summary statement of the principles of their preservation strategies with links to the more detailed documents, including high level requirements, standards, and other resources. The University of Illinois Urbana-Champaign provides a useful model:

- Development and maintenance of reliable options for the ingest of new materials into the repository, based on community standards or best practices;
- Provision of reliable data management services for timely access to deposited content;
- Development and maintenance of archival storage for deposited content;
- Conducting IDEALS management and administrative activities in such a manner as to further the program's mission of preserving deposited content;
- Monitoring and remaining active in community preservation activities, best practices and standard; and
- Developing local preservation planning activities that will anticipate and respond to changes in the preservation environment (e.g. format migration or emulation strategies).

Other approaches may include listing the specific formats that are supported.

5 Metadata Creation

A simple statement of policy regarding metadata creation and maintenance is sufficient. For example, Yale's plan states: "Metadata is fundamental to preserving Yale University Library's digital resources. Preservation metadata includes a number of different types of metadata…Particular attention is paid to the documentation of digital provenance…and the relationship among different objects within preservation repositories." Such a policy statement can be included as a separate section or incorporated under the organizational commitment section.

CONTINUED »

Due to the ongoing changes in the digital environment, it is important that regular monitoring of the environment is incorporated into the digital plan.

6 Roles and Responsibilities

Like metadata creation, roles and responsibilities can be a separate section or included under organizational commitment. Individual institutions in particular may wish to incorporate it under the organization's commitment, where the roles and responsibilities can be very detailed. The University of Kansas plan, for example, provides very detailed descriptions. Digital preservation programs operated by collaboratives may wish to have a separate responsibilities section.

7 Training

OAIS requires that continuing development of staff be addressed in the digital preservation program, however few organizations provide sufficient support for staff development. The University of Kansas Library Preservation Planning for Digital Information lays out an institution-wide strategy for training with the following principle: "Key to the success of digital preservation planning on the University of Kansas campus is the recruitment and involvement of staff at all levels of the University." The suggested curriculum focuses on five areas: general awareness, information lifecycle management, information storage management and systems, maintenance, best practices and standards, and legal issues and university policies.

8 Monitoring and Review

Due to the ongoing changes in the digital environment, it is important that regular monitoring of the environment is incorporated into the digital plan. This monitoring is an OAIS requirement and is broadly defined. Each preservation program will need to consider its particular "environment." Clearly the technical environment will require monitoring and there are national and international efforts to facilitate that monitoring. However, programs will need to additionally monitor the legal environment that includes international, national, state, local, and institution legislation and procedures. Regulations and statutes can change as rapidly as technology and may be more difficult to monitor. Additionally, the organization's political environment requires monitoring; the more closely the mandate for the digital program is tied to the political environment the more closely that environment may need to be monitored. The business environment may also require monitoring; mergers and acquisitions may impact support from vendors and other partners. And it is especially critical to monitor the status of partners; key staff and funding changes and funding changes at partner organizations may necessitate a program review.

As the environment changes, plans need to be revised and updated. The digital preservation plan should be reviewed annually. Responsibility for the review should be clearly established, along with the review procedure and the required timeframe of the review. | FE | doi: 10.3789/isqv22n2.2010.02

LIZ BISHOFF <Liz.Bishoff@gmail.com> is head of The Bishoff Group, LLC and was formerly Director, Digital and Preservation Services at the Bibliographical Center for Research (BCR).

Sample Plans and Policies

A written preservation policy demonstrates an organization's commitment to digital preservation. The resources below offer insight into the reasons for having a preservation plan, factors to consider in developing a plan, and the areas the plan should address. They also provide several examples of existing digital preservation policies.

Au Yeung, Tim. *Digital Preservation: Best Practice for Museums*. Commissioned by the Canadian Heritage Information Network. Gatineau, Quebec, Canada: Minister of Public Works and Government Services, 2004. http://www.pro.rcip-chin.gc.ca/sommaire-summary/preservation_numerique-digital_preservation-eng.jsp

Au Yeung, Tim. *Digital Preservation for Museums: Recommendations*. Commissioned by the Canadian Heritage Information Network. Gatineau, Quebec, Canada: Minister of Public Works and Government Services, 2004. http://www.pro.rcip-chin.gc.ca/sommaire-summary/preservation_recommandations-preservation_recommendations-eng.jsp

Columbia University Libraries. *Policy for Preservation of Digital Resources*, July 2000, revised 2006. http://www.columbia.edu/cu/lweb/services/preservation/dlpolicy.html

Cornell University Library. *Cornell University Library Digital Preservation Policy Framework*, December 2004. http://commondepository.library.cornell.edu/cul-dp-framework.pdf

ERPA Guidance: Digital Preservation Policy Tool. Electronic Resource Preservation and Access Network, September 2003. http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

Florida Center for Library Automation. *Florida Digital Archive (FDA) Policy Guide*, version 2.5, April 2009. http://www.fcla.edu/digitalArchive/pdfs/FDAPolicyGuideversion2.5.pdf

Fyffe, Richard et al. *Preservation Planning for Digital Information*. Lawrence, KS: University of Kansas, November 11, 2004. https://kuscholarworks.ku.edu/dspace/bitstream/1808/166/1/Preservation%20Planning%20for%20Digital%20Information.pdf

MetaArchive Cooperative. *MetaArchive Cooperative Charter*, November 2009. http://www.metaarchive.org/public/resources/charter_member/MetaArchive_Charter_2010.pdf

MetaArchive Cooperative. "Appendix B: Membership Agreement," in *MetaArchive Cooperative Charter*, November 2009. http://www.metaarchive.org/public/resources/charter_member/Membership_Agreement_2010.pdf

McGovern, Nancy Y. *Digital Preservation Policy Framework: Outline*, version 2.0, January 2007, revised October 2007. http://www.icpsr.umich.edu/files/ICPSR/curation/preservation/policies/dp-policy-outline.pdf

McGovern, Nancy Y. ICPSR *Digital Preservation Policy Framework*, November 2007. http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/preservation/policies/dpp-framework.jsp

National Digital Information Infrastructure and Information Preservation Program, Library of Congress. *Sustainability for Digital Formats: Planning for Library of Congress Collections*. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

National Library of Australia. *Digital Preservation Policy*, 3rd ed., 2008. http://www.nla.gov.au/policy/digpres.html

Schreibman, Susan, ed. *Best Practice Guidelines for Digital Collections at University of Maryland Libraries*, 2nd ed., May 4, 2007. http://www.lib.umd.edu/dcr/publications/best_practice.pdf

University of Illinois at Urbana-Champaign. *IDEALS Digital Preservation Policy*, November 2009. https://services.ideals.uiuc.edu/wiki/bin/view/IDEALS/IDEALSDigitalPreservationPolicy

Yale University Library. *Policy for Digital Preservation*, November 2005, revised February 2007. http://www.library.yale.edu/iac/DPC/revpolicy2-19-07.pdf

RELEVANT LINKS

Columbia University Libraries. Policy for Preservation of Digital Resources, July 2000, revised 2006.
www.columbia.edu/cu/lweb/services/preservation/dlpolicy.html

Fyffe, Richard et al. Preservation Planning for Digital Information. Lawrence, KS: University of Kansas, November 11, 2004.
https://kuscholarworks.ku.edu/dspace/bitstream/1808/166/1/Preservation%20Planning%20for%20Digital%20Information.pdf

McGovern, Nancy Y. ICPSR Digital Preservation Policy Framework. November 2007.
www.icpsr.umich.edu/icpsrweb/ICPSR/curation/preservation/policies/dpp-framework.jsp

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

Trusted Digital Repositories: Attributes and Responsibilities. Mountain View, CA: Research Libraries Group, May 2002.
www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf

University of Illinois at Urbana-Champaign. IDEALS Digital Preservation Policy. November 2009.
https://services.ideals.uiuc.edu/wiki/bin/view/IDEALS/IDEALSDigitalPreservationPolicy

Yale University Library. Policy for Digital Preservation, November 2005, revised February 2007.
www.library.yale.edu/iac/DPC/revpolicy2-19-07.pdf

THE UNIFIED DIGITAL FORMATS REGISTRY

ANDREA GOETHALS

Why do we need a format registry for digital preservation?

If you diligently protected a WordStar document for the last twenty-five years, all of its original bits may still be intact, but it would not be usable to anyone. Today's computers do not have software that can open documents in the WordStar format. It's not enough to keep digital bits safe; to fully preserve digital content we must make sure that it remains compatible with modern technology. Given that the ultimate goal of digital preservation is to keep content usable, practically how do we accomplish this? Somehow we need to be able to answer two questions: (1) is the content I'm managing in danger of becoming unusable, and if so, (2) how can I remedy this situation?

Formats play a key role in determining if digital material is usable. While traditional books are human-readable, giving the reader immediate access to the intellectual content, to use a digital book, the reader needs hardware that runs software, that understands formats, composed of bits, to access the intellectual content. Without technological mediation, a digital book cannot be read. Formats are the bridge between the bits and the technologies needed to make sense of the bits. The formats of the bits are the key to knowing if there are technologies that can make the bits usable.

Returning to the question—Is the content I'm managing in danger of becoming unusable?—the question can be answered if we know the formats of the content we're managing, and additional information about those formats. We

CONTINUED »

Formats are the bridge between the bits and the technologies needed to make sense of the bits.

CONTINUED »

need to know if there are current acceptable technologies that support the formats, sustainability issues related to the formats, and how others in the digital preservation community have assessed the formats. If we determine that the content is in danger of becoming unusable, we can form a remediation plan if we have additional information about the formats. We need to know alternative formats for the content, supporting transformation or emulation tools, and as a last resort, enough documentation about the format to construct our own tools to transform or render the content.

All institutions engaged in long-term digital preservation need this same format information. The concept of the format registry is simple—pool and share the data so that each institution does not have to collect and manage this information for itself, and does not need in-house expertise for all the formats it needs to manage. Additionally, because the format registry would provide authority control for format names and identifiers, it would enable institutions to more easily share file tools and services, and exchange content.

History of the format registry initiative

The first planning sessions for what came to be known as the Global Digital Format Registry, or GDFR, were sponsored by the Digital Library Federation (DLF) in 2003. These meetings were attended by policymakers and technologists from various national libraries and archives, academic research libraries, universities, library organizations, and standards bodies. Out of these meetings came a clear rationale for a shared format registry, over thirty use cases demonstrating how the registry could be used in preservation operations, and preliminary designs.

Following those meetings, Harvard University agreed to seek funding for and host the first instance of the registry. The Mellon Foundation funded a two year project beginning in 2006, and the development was subcontracted out to OCLC. The project produced a very detailed data model, and a registry model based on shared governance, cooperative data contribution, and distributed data hosting. When the project ended in 2008, a proof of concept registry at Harvard containing a limited amount of format information was made available on the Internet.

Following the project, Harvard began to plan next steps for the registry. The proof of concept registry would need additional technical work to turn it into a full-fledged registry. In addition, there were a number of governance issues still to be resolved to make the registry sustainable. It would need long-term administrative, operational, and financial resources. The reality, however, was that the registry landscape had changed a great deal from when the GDFR project began. Now there was already in existence another format registry that was being used by many in the preservation community: PRONOM.

PRONOM, developed by The UK National Archives (TNA), was created to meet TNA's requirements, but the registry information was freely shared on the Internet. Like the GDFR, PRONOM contains information about formats as well as related software, hardware, media, documents, and organizations. It's not a coincidence that the GDFR and PRONOM data models are similar. TNA was a significant contributor to the GDFR effort and the GDFR and PRONOM teams shared data model information so that they would be compatible. The intention was that PRONOM would become a node in the GDFR network of format registries when GDFR became fully operational. However, in 2008 when Harvard started to look at next steps for the GDFR, it was clear that PRONOM was further along technologically and in terms of use by the preservation community. But because PRONOM is owned and maintained by a single institution, it was not possible for other institutions to contribute information to the registry, and the community had become reliant on a single institution for sustaining an essential piece of preservation infrastructure.

This was the dilemma: neither GDFR nor PRONOM alone was fulfilling the long-term requirements for the digital preservation community. The community needed the format information and services already provided by PRONOM but also wanted the shared governance, cooperative data contribution, and distributed data hosting promised by GDFR.

Progress: UDFR established

In early 2009, the National Archives and Records Administration (NARA) hosted a format registry planning meeting, which included members of the GDFR and PRONOM teams. In this meeting it was agreed that it would be advantageous for all to combine the PRONOM and GDFR initiatives into a single registry—the Unified Digital Format Registry. UDFR would include the services and data of PRONOM and support the shared governance, cooperative data contribution, and distributed data hosting of GDFR.

The work required to establish the UDFR falls into two general categories: governance and technical work. The governance work includes designing and implementing the plan for ongoing UDFR governance, funding, and operations. The technical work includes the design, development, and testing of registry software and processes needed to exchange registry information with tools, services, and repositories. To address this work, an interim governing body and a technical working group were formed consisting of members from national and academic libraries, universities, and national archives who had participated in the earlier registry initiatives. These groups formed a plan to quickly put into place an operational first version of the registry, while working in parallel to replace the interim governance body with a permanent governance structure for UDFR.

Working from documents created for the GDFR and PRONOM projects, the technical working group compiled the requirements that should be implemented in the first version of the UDFR:

- A publicly accessible web-based user interface that can be used to search, browse, display, and download registry records
- An API for tools and services to query, retrieve, and export registry records for use in local repositories or applications
- Ability to export information to DROID, a format identification tool created by TNA
- Automatic tracking of the history of registry information changes
- Population of the registry with all of the PRONOM content

Near the end of 2009, the governance working group submitted a proposal to the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) to fund the one-year program of technical work needed to establish the first version of the UDFR. Under the proposal the work would be conducted at the University of California Curation Center (UC3) of the California Digital Library (CDL). UC3 will provide project oversight and management and will hire two new staff for the project—a project architect and a developer. The proposal was accepted by the Library of Congress in early 2010 and UC3 has now begun the hiring process for the project, which is scheduled to run from July 2010 to July 2011.

Future plans: UDFR and beyond

In parallel to the technical work that will occur at the UC3 over the next year, the interim governance working group will establish the permanent governing body for the UDFR. This permanent group is needed to define registry policies and procedures, such as the editorial process to ensure registry information is accurate, how future enhancements will be defined and prioritized, and intellectual property policies related to the registry software and information. In addition, this group is needed to fund the UDFR's administration, maintenance, and future enhancements.

A key future enhancement is to transform the initial UDFR design into a network. Initially there will be a single registry instance hosted by UC3. However, the long-term goal of the UDFR project is to establish a network of registry instances operated by various institutions around the world, with automatic processes to copy the UDFR content among the registry instances. This will increase the safety of the registry information and reduce the dependency on any single institution.

The initial version of the UDFR will provide interoperability with existing applications used for digital preservation. It will supply format identification information to DROID, and it will provide export services that could be used to import format or environment information into local repository databases. Ultimately though, it is the intention that the UDFR will serve as a source of format information to many tools and services that will be developed by the preservation community over time for format identification, assessment, validation, characterization, transformation, delivery, and emulation. | FE |

doi: 10.3789/isqv22n2.2010.04

ANDREA GOETHALS <andrea_goethals@harvard.edu> is Digital Preservation and Repository Services Manager at Harvard University Library.

RELEVANT LINKS

DROID
sourceforge.net/projects/droid/

Global Digital Format Registry
www.gdfr.info/

NDIIPP
www.digitalpreservation.gov/library/

PRONOM
www.nationalarchives.gov.uk/PRONOM/

Unified Digital Format Registry
udfr.org

University of California Curation Center (UC3)
www.cdlib.org/services/uc3/

Selecting Formats for Digital Preservation:

Lessons Learned during the Archivematica Project

EVELYN PETERS MCLELLAN

The Archivematica project was launched a year ago as a follow-up to a report entitled *Towards an Open Source Repository and Preservation System* which was released by the UNESCO Memory of the World Sub-Committee on Technology in 2007. The report surveyed then existing open-source tools for digital preservation, concluding that although a wide variety of such tools were available, they were neither comprehensive nor integrated enough to provide a complete ingest-to-access environment for preserving digital records.

The report recommended that UNESCO support "the aggregation and development of an open source archival system, building on and drawing together existing open source programs." This is the goal of Archivematica, a preservation system being developed with funding from UNESCO and in collaboration with the City of Vancouver Archives. Software development has thus far been led by Artefactual Systems, a Canadian open-source software company specializing in information management systems for archives and libraries.

Archivematica is an integrated environment of open-source tools which can be deployed from a single installation on any operating system. (The system is based on Ubuntu Linux but can be implemented in a virtualized environment, allowing it to be run on top of any number of host operating systems such as Microsoft Windows.) The software and the source code are freely available online under a GPL license. The system is based on a detailed use case analysis of the ISO Open archival information system (OAIS) functional model and supports best practice metadata standards such as PREMIS, METS, Dublin Core, EAD, and ISAD(G). Detailed workflow documentation assists the user to move an Information Package through submission, integrity checking, identification and validation, normalization, packaging into an Archival Information Package, storage, and provision of access. One of the key components of this process is normalization, the conversion of digital objects into a small number of standard preservation-friendly formats on ingest. This is the part of the Preservation Planning function of OAIS which "receives archive approved standards and migration goals from Administration," including "format standards," the goal being to ensure that the preserved objects remain accessible and usable in the future despite issues of technological obsolescence and incompatibility.

Archivematica supports emulation preservation plans by preserving original bitstreams, and it supports migration preservation plans by monitoring at-risk file formats and providing a process to migrate them at a future date. Nevertheless, Archivematica's default preservation strategy is to normalize digital objects into preservation formats upon ingest in order to make best use of the limited time that organizations will have to process and monitor large, diverse collections of digital objects.

Building normalization paths into the software requires choosing target formats and integrating open-source tools to perform the migrations. The choice of formats is based on four basic criteria which will be familiar to many of those who have experience with digital preservation:

1. The specification must be freely available.
2. There must be no patents or licenses on the format.
3. Other established digital repositories should be using or have endorsed the format.
4. There should be a variety of writing and rendering tools available for the format.

Selection of formats has been an iterative process of researching best practices, testing normalization tools, and, as far as possible, comparing before and after results of conversions by measuring significant properties. During this process it was found that selecting target formats based on the first three criteria is not difficult, since a great deal of research has been done on the subject and de facto standards have been proposed and in some cases implemented. However, there are some significant challenges with the fourth criterion. Specifically, for this project there need to be open-source tools available for conversion from original formats. This is to ensure that the tools can be integrated and distributed with the existing tools in the system, which must remain entirely free of software license restrictions and costs. Another important consideration is that they offer a Linux command-line interface to enable full ingest process automation. Thus far, it has been the Archivematica team's experience that the scarcity of some types of tools and inadequacy of others has made the process of selection considerably more difficult, illustrating the challenges that can arise when moving from the realm of the ideal to the realm of the practical.

Moving image files provide an example of some of the difficulties involved. A consensus seems to be building in the research community that Motion JPEG2000 is the desired target format because it provides mathematically lossless, wavelet compression. (See, for example, *Lossless Video Compression for Archives: Motion JPEG2k and Other Options*.) Motion JPEG2000 was adopted as an ISO standard in 2001 (ISO/IEC 15444-3); however, during the nine years of its existence only a handful of tools have emerged to convert to it, and these are proprietary and not designed for use with Linux. Most heritage institutions that are converting to Motion JPEG2000 are converting directly from analog video using specialized hardware. There are a number of open-source Linux-based tools for converting moving image files from one digital format to another, most notably FFmpeg and Avidemux, but they do not currently encode to Motion JPEG2000. For these reasons, the default normalization path for moving image files in Archivematica is MPEG-2. MPEG-2 is a reasonably well-accepted preservation format; for example, the Library of Congress (with some reservations and qualifications) and Library and Archives Canada both recommend it. However, MPEG-2 compression is not entirely lossless. If and when an appropriate Motion JPEG2000 normalization tool becomes available, it will be added to Archivematica and users will then have the option to migrate the original moving images to Motion JPEG2000 and discard the existing MPEG-2 versions.

Even more problematic are Microsoft Office files. Theoretically, it is a simple task to choose the XML-based Open Document Format (ODF) and PDF/Archival (PDF/A), both ISO standards, as preservation formats. (The Archivematica team briefly considered using Office Open XML, the Microsoft XML format that was approved as an ISO Standard in 2008. However, there are no open-source tools that convert to the format at present, and because at over 6,000 pages the standard

Specifically, for this project there need to be open-source tools available for conversion from original formats. This is to ensure that the tools can be integrated and distributed with the existing tools in the system, which must remain entirely free of software license restrictions and costs.

CONTINUED »

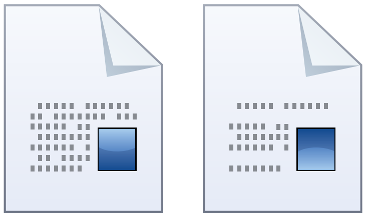
is so complex and lengthy it is unlikely that any such tools will emerge in the near future.) The user could choose to use one or the other of ODF or PDF/A or both. There are Linux-based tools to convert to ODF, including Xena, the National Archives of Australia's bulk normalization tool. To normalize office documents, Xena calls on OpenOffice to manage the conversion of a Microsoft Office document to ODF. Unfortunately, because OpenOffice has had to rely on reverse engineering of proprietary Microsoft specifications to map to ODF, the formatting of the converted document often differs from the original; the differences may be minor but they can change the overall look and feel of the document, which may call into question the authenticity of the conversion.

This problem becomes even more critical with PDF/A, since one of the most compelling reasons for using that format is to provide an accurate visual representation of the original. There are very few open-source bulk normalization tools to convert to PDF/A, and those that do (such as OpenOffice) must, once again, rely on reverse engineering of closed specifications in order to perform the conversion. When OpenOffice opens a Microsoft document, the document renders with some changes to formatting; the PDF/A is created from this altered rendering.

Conversion using a plug-in that works directly with the native application is the most direct path to success with either ODF or PDF/A. The proprietary Adobe Acrobat Distiller, for example, works directly with Microsoft software to produce visually true conversions to PDF/A. Similarly, Sun Microsystems has produced a free (but not open-source) plug-in to convert documents to ODF from within Microsoft applications. However, the problem remains that there is no way to integrate these tools into any freely available open-source digital preservation system. Following the example set by the National Archives of Australia, the Archivematica team has chosen to build default normalization to ODF into its system and is currently testing conversions to determine which tool to use. As with moving image files, acceptable conversions might not be possible for the immediate future and bulk Microsoft Office migration processes may need to be run at a later time when better tools become available. PDF/A remains under consideration as a preservation format, but more time is needed to evaluate available tools before it can be built into the system as a default.

Fortunately, some types of files lend themselves more easily to normalization. Numerous raster image formats, for example, can be converted easily to uncompressed TIFF 6.0 using ImageMagick, and FFmpeg does a good job of converting audio files to uncompressed (LPCM) WAVE files. Both of these are well accepted preservation formats in the library and archives community. However, the lack of tools for other kinds of digital objects means that, with regard to normalization, any open-source integration of digital preservation tools must remain a work in progress. The driving goal behind the Archivematica project has been to lower the barriers to digital preservation for institutions which may have limited technical and financial resources. The best way to do this is to provide a complete system that can be freely downloaded, used, distributed, and modified by any individual or institution. At this time, Archivematica incorporates best-possible normalization paths, and the team has adopted an agile development process in which the system incorporates new tools as soon as they become available. This is the most effective way to work within current limitations, and it is the most realistic means of achieving UNESCO's goal of bringing open-source digital preservation capability to institutions all over the world. | IP |

doi: 10.3789/isqv22n2.2010.05



Because OpenOffice has had to rely on reverse engineering of proprietary Microsoft specifications to map to ODF, the formatting of the converted document often differs from the original; the differences may be minor but they can change the overall look and feel of the document, which may call into question the authenticity of the conversion.

The driving goal behind the Archivematica project has been to lower the barriers to digital preservation for institutions which may have limited technical and financial resources. The best way to do this is to provide a complete system that can be freely downloaded, used, distributed, and modified by any individual or institution.

EVELYN PETERS MCLELLAN <evelyn@artefactual.com> is Systems Archivist at Artefactual Systems Inc., Vancouver, Canada.

RELEVANT LINKS

Archivematica project
www.archivematica.org

Archivematica Media Type Preservation Plans
www.archivematica.org/wiki/index.php?title=Media_type_preservation_plans

Avidemux
fixounet.free.fr/avidemux/

Bradley, Kevin, et al. Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open source Digital Archival and Preservation System and on Related Software Development. UNESCO Memory of the World Sub-Committee on Technology, June, 2007.
portal.unesco.org/ci/en/ev.php-URL_ID=24700&URL_DO=DO_TOPIC&URL_SECTION=201.html

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

FFmpeg
ffmpeg.org/

Gilmour, Ian and R. Justin Dávila. Lossless Video Compression for Archives: Motion JPEG2k and Other Options. Media Matters, January 2006.
www.media-matters.net/docs/WhitePapers/WPMJ2k.pdf

Guidelines for Computer File Types, Interchange Formats and Information Standards. Library and Archives Canada, 2004.
www.collectionscanada.gc.ca/government/products-services/007002-3017-e.html

ImageMagick software
www.imagemagick.org/

Information technology – JPEG 2000 image coding system: Motion JPEG 2000, ISO/IEC 15444-3:2007
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41570

Information technology – Open Document Format for Office Applications (OpenDocument) v1.0, ISO/IEC 26300:2006
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485

Material Exchange Format (MXF), SMPTE 377-1-2009
en.wikipedia.org/wiki/MXF

MPEG-2, Generic coding of moving pictures and associated audio information, ISO/IEC 13818 (9 parts)
mpeg.chiariglione.org/standards/mpeg-2/mpeg-2.htm

MPEG-2 Video Encoding (H.262). In: Sustainability of Digital Formats: Planning for Library of Congress Collections.
www.digitalpreservation.gov/formats/fdd/fdd000028.shtml

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

Xena software
xena.sourceforge.net/

Format Considerations in Audio-Visual Preservation Reformatting:

Snapshots from the Federal Agencies Digitization Guidelines Initiative

CARL FLEISCHHAUER

Introduction

Digitization practices have developed and matured in phases. Documents, books, and photographs were among the first items to be digitized by memory institutions—roughly speaking, beginning in the 1980s—and the practices for making still images from these source materials are reasonably mature. The digitization of sound recordings made headway in the late 1990s, with the last decade bringing good levels of consensus on the best approaches to use. Although mature, however, the practices for creating still images and digital audio continue to be refined. Meanwhile, practices for the preservation digitization of moving image content—at least in our memory institutions—are still in their infancy.

Using examples from the Federal Agencies Digitization Guidelines Initiative, this article will provide a few snapshots of digital reformatting practices with an emphasis on formats as they continue to evolve and, for moving images, as they begin to emerge. The federal agencies initiative has two Working Groups. The Still Image Working Group is concerned with the reformatting of books, manuscripts, photographs, maps, and the like, while the Audio-Visual Working Group is concerned with sound recordings, video recordings, and motion picture film. This writer coordinates the Audio-Visual Working Group and the description that follows concerns recorded sound reformatting (with a glimpse of the still image environment) and the group's exploration of moving image content.

"What formats do you recommend?" That is a question we often hear and, more often than not, people expect a three-letter answer, e.g., *wav, mpg,* or *mxf.* Alas, just naming a file format only begins to answer the question. In addition to the file format as container—what the three letters point to—we must attend to the encoding of the data within the container, its organization, and its internal description. My use of the terms *format* and *formatting* is in sync with the usage of the Library of Congress Format Sustainability website. (See the *What is a Format* page.)

The work of the Federal Agencies Working Groups is currently focused on *files*. All reformatting activities produce files and this common ground makes a good fit for interagency deliberations. Members of both Working Groups, to be sure, understand the importance of digital resources comprised of multiple files: *packages* in the parlance of the Open Archival Information System. Searchable access to digital resources is often provided at the package level. In a library setting, *packages* often correlate to what are called *manifestations* in the terminology of the Functional Requirements for Bibliographic Records (FRBR). Library cataloging typically describes content manifestations. In an archive, digital *packages* generally correlate to an item in, say, an EAD (Encoded Archival Description) finding aid, where items are typically part of series and collections or record groups. However, the practices for packaging digital resources vary so much from agency to agency (and even within agencies) that we decided "files first, packages later."

In our considerations, three aspects of formatting are at stake:

1. The **file format**, what is sometimes called the *container* for the encoded bitstreams and other elements

2. The **encoded bitstream**, i.e., the content *data*, what is often called the *essence* in broadcast and professional media production circles

3. The **metadata** that is embedded in the file, inevitably including some technical metadata ("you can't open a file in an application without it"), sometimes supplemented by judiciously chosen elements of descriptive and administrative metadata

The practices for packaging digital resources vary so much from agency to agency (and even within agencies) that we decided "files first, packages later."

Embedded metadata

Most archives and libraries that manage digital content depend upon the metadata in databases, integrated library systems, and/or digital content management systems. These systems or their extensions also support patron discovery and retrieval of digital content. Thus we all tend to think of these database and database-like systems as the real home for our metadata, although they generally do not include the finest-grained elements of technical information about the content, e.g., the color space of an image file.

What is the value, then, of file-embedded metadata? The charter for the Federal Agencies subgroup devoted to the topic states that embedded metadata plays an important role "in the management, use, and sustainability of digital assets," noting that the adoption of practices that take advantage of such metadata have been inhibited by "the lack of clear, comprehensive, and uniform guidelines." The preservation-related importance of embedded metadata is also expressed in one of the Working Group's use cases for archival master images: "Disaster recovery in the event of the impairment of digital asset management systems depends upon the availability of metadata in standardized formats, including embedded image-level metadata and work-level descriptive, administrative, and structural metadata." Meanwhile, at the Format Sustainability website, *self documentation*, which refers to embedded metadata, is defined as one of the sustainability factors for digital formats.

Beyond reformatting, embedded metadata takes on special importance for libraries or archives that receive born-digital content. The acquisition of digital content with a useful mix of descriptive, administrative, and technical metadata in standardized structures will reduce the effort required to ingest and manage that content over the long term. Leaving long term management aside, it is fair to say that the seemingly simple action of transferring digital content from one organization to another is well supported by the presence of embedded metadata.

The Library's interest in promoting the embedding of metadata by content creators accounts for our support of efforts like PhotoMetadata.org, organized by the Stock Artists Alliance. We endorse the idea of embedding at least some metadata at or near the beginning of the content lifecycle. The PhotoMetadata outreach activity received matching funds from the Library's National Digital Information Infrastructure and Preservation Program (NDIIPP), and it encourages photographers to make good use of the metadata specifications from the International Press Telecommunications Council (IPTC), as a supplement to the EXIF metadata (a standard of the Japan Electronics and Information Technology Industries Association, JEITA) that is embedded in files by the camera.

CONTINUED »

The federal agencies initiative has two Working Groups. The Still Image Working Group is concerned with the reformatting of books, manuscripts, photographs, maps, and the like, while the Audio-Visual Working Group is concerned with sound recordings, video recordings, and motion picture film.

One sign of the maturity of recorded sound reformatting practices was the absence of debate within the Working Group about file formats and bitstream encoding. Every participating audio specialist accepted the idea that the file format should be WAVE and that the encoding should take the form of linear pulse code modulation (LPCM).

WAVE files for recorded sound

One sign of the maturity of recorded sound reformatting practices was the absence of debate within the Working Group about file formats and bitstream encoding. Every participating audio specialist accepted the idea that the file format should be WAVE (more on this in a moment) and that the encoding should take the form of linear pulse code modulation (LPCM). This consensus owes a great debt to the work carried out over the last decade by the International Association of Sound and Audiovisual Archives (IASA) and to pathfinding projects like *Sound Directions*, carried out at Indiana and Harvard Universities.

Sound quality correlates to the sampling frequency and bit depth selected for LPCM encoding. Both IASA and Sound Directions push for sampling rates of 96 kilohertz (with a bit of grudging room for 48) and a bit depth of 24 per sample. For comparison, audio compact disks are pegged at 44.1 kilohertz and 16 bits per sample, considered to be inferior for archival masters. Members of the Working Group concur in these judgments.

The name WAVE is generally glossed as short for *waveform audio format*. The file format is one of the subtypes of the more generic RIFF (Resource Interchange File Format) format, whose specification was published in 1991 by Microsoft and IBM to serve the then-new Windows 3.1 operating system. In turn, WAVE has its own subtypes, one of which is especially important to the Working Group: the Broadcast WAVE Audio File Format (nicknamed BWF or BWAV), developed in the late 1990s by the European Broadcast Union (EBU).

Although WAVE was created in the private sector, the relevant specifications are publicly available and, as noted, the format has formed the basis for additional work by the EBU standards body. (In this aspect, WAVE can be compared to TIFF, usually glossed as *Tagged Image File Format*, an open proprietary specification, now from Adobe, that has provided the foundation for ISO standardization efforts like TIFF/EP and TIFF/IT.) WAVE and its RIFF siblings have several virtues, including that their architecture is transparent and they can be written and read in a number of software applications.

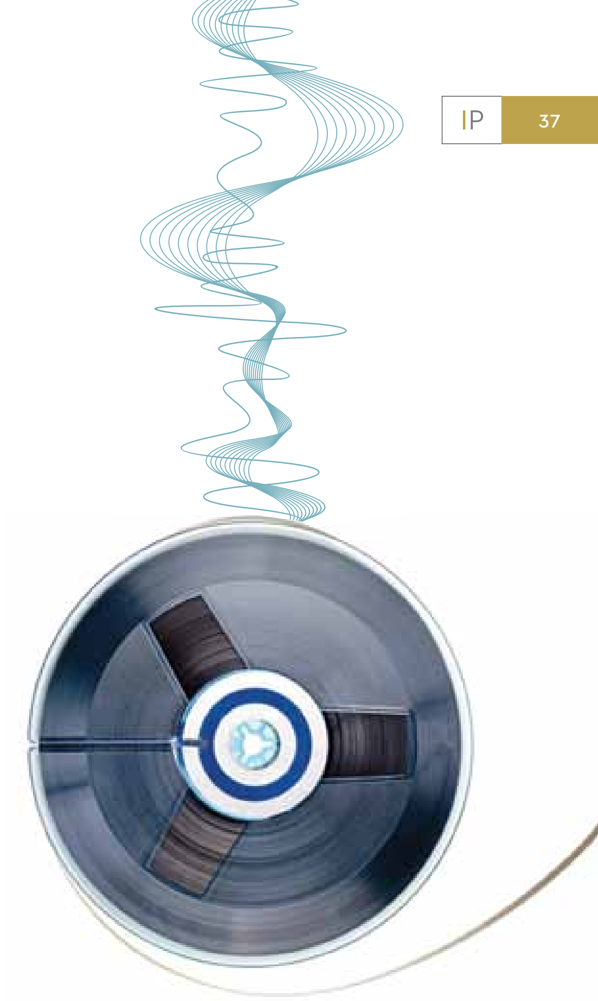
The underlying structure for the RIFF format family consists of what are called *chunks*. The specification permits anyone to add new chunks, which is exactly what the EBU did when it specified the BWF format. Applications that play or read RIFF-family files are designed to harmlessly skip over chunks they do not understand. The structural transparency of formats like WAVE and the BWF subtype together with their widespread adoption—they are readable in many applications—make them very sustainable choices for the preservation of recorded sound.

WAVE files employ 32-bit addressing and this limits their size to 4 gigabytes (2 GB in some software applications or operating systems). Many recordists today produce high resolution files that exceed these limits and that has led to extended specifications for WAVE and BWF files. These new formats are closely patterned on their predecessors but they employ 64-bit addressing. This permits files of virtually any size, up to the limits of available disk space on a given workstation. The extended documentation includes a Microsoft specification referred to as WAVEFORMATEXTENSIBLE and the EBU standard *An Extended File Format for Audio* (EBU-TECH-3306-2007). For the time being, the Working Group is deferring an examination of 64-bit extended formats.

Metadata in WAVE files

Although happy to minimize the discussion of audio file formats and encodings, the Working Group spent some time refining a guideline for embedding descriptive and administrative metadata in WAVE files. The Working Group saw no need for action regarding the technical file-characteristics metadata required by playback applications in order to open a given file. This type of metadata is provided by the *format chunk* defined by the 1991 Microsoft-IBM RIFF specification. (The actual essence bitstream is contained in the RIFF data chunk; in the case of a WAVE file, this is the recorded sound data.) Additional information on these chunks will be found in an explanatory paper from the Working Group: *Embedding Metadata in Digital Audio Files*.

Existing WAVE specifications define some chunks for descriptive and administrative metadata. The 1991 Microsoft-IBM RIFF specification defines the *LIST info chunk*, more often referred to as the *INFO chunk*, which includes twenty-odd tagged elements ranging from *title* to *copyright* to *dots per inch* (for an image file). As far as we were able to determine,



the INFO chunk (or family of subchunks) is typically used by practitioners (not archivists) in fairly loose fashion.

Meanwhile, the BWF specification family adds three metadata chunks to WAVE: the widely adopted *bext chunk* (formally the *broadcast extension*) and the less widely used *aXML chunk* and *iXML chunk*. aXML is named after an XML expression of the Dublin Core-based core audio descriptive metadata standard. The specification allows for the storage of any valid XML document (version 1 or higher) that may be of any length (limited by RIFF specifications) and may appear in any order with the other chunks. The aXML chunk does not constrain how the user defines the data. The iXML chunk was created by audio hardware and software manufacturers to facilitate transfer of production metadata across systems. The chunk contains a defined XML document for production information such as *project*, *tape*, *note*, and *user*. On paper, the aXML and iXML chunks have much to recommend them, including an XML approach and a relatively large capacity for data. The lack of adoption and the consequent shortage of tools for writing and reading data to those chunks, however, led the Working Group to set aXML and iXML aside for now and to concentrate on making the most of the bext chunk.

The BWF bext chunk offers nine elements, generally constrained by low character counts, and customarily inscribed as ASCII strings. One of the nice touches is an element named *CodingHistory*, in which you can write a very short story about where the sound came from and how it was transferred. Here's an example (and a translation) of CodingHistory:

```
A=ANALOG,M=mono,T=Studer816; SN1007; 15 ips; open reel tape,    1
A=PCM,F=96000,W=24,M=mono,T=Pyramix1; SN16986,                  2
A=PCM,F=96000,W=24,M=mono,T=Lynx; AES16; DIO,                   3
```

Explanation: Line 1 reads: an analog, mono, open-reel tape played back on a Studer 816 tape machine with serial number 1007 at tape speed 15 ips. **Line 2 reads:** tape was digitized to PCM coding in mono mode at 96 kHz sampling frequency and 24 bits per sample on a Pyramix 1 DAW with serial number 16986. **Line 3 reads:** the audio was stored as a BWF file with PCM coding in mono mode at 96 kHz sampling frequency and 24 bits per sample using a Lynx AES16 digital input/output interface.

As the example indicates, CodingHistory does not permit the elaborate descriptions that are possible with the extension schemas typically used in METS (Metadata Encoding and Transmission Standard) implementations, under the headings *sourceMD* (about the item you started with) and *digiprovMD* ("digital provenance," about the conversion process you used when reformatting). Two very rich schemas that make great candidates for METS extensions have been defined by the Audio Engineering Society, usually referred to as *Administrative and Structural Metadata for Audio Objects* and *Process and Handling History of Audio*. Draft versions of these standards were employed in the Sound Directions project. The Working Group is not aware of any practice that embeds this metadata in files, although presumably the EBU aXML chunk could be used in this way.

As the example indicates, CodingHistory does not permit the elaborate descriptions that are possible with the extension schemas typically used in METS (Metadata Encoding and Transmission Standard) implementations, under the headings *sourceMD* (about the item you started with) and *digiprovMD* ("digital provenance," about the conversion process you used when reformatting).

As we drafted our WAVE guideline, we were repeatedly struck by the relatively skeletal nature of the bext chunk and the imperfectly-defined INFO list chunk. To

CONTINUED »

When promoting a guideline or standard, one of the issues to address concerns the ease with which the user community can comply: are there tools for the job? After publishing our guideline for metadata in the EBU bext chunk and the RIFF/WAVE INFO chunk, we asked our expert consultants from AudioVisual Preservation Solutions to produce an edit-and-embedding tool.

be fair, this state of affairs is understandable: the bext chunk had been designed to support the exchange of program content between broadcasters, using only a few data elements written in short ASCII strings. The INFO list was designed in the early days of digital formatting, before practitioners had sophisticated views of content transfer and identification.

Identifiers were a point of particular concern for the Working Group. The bext specification defines one main element for an identifier (*OriginatorReference*) and it is limited to 32 characters. (In the second version of the specification, there was also a place defined for the Unique Material Identifier (UMID) defined by the Society of Motion Picture and Television Engineers as standard 330M.) In the reformatting work carried out by our member agencies, there is often an interest in recording an identifier for "the original" and another for the digital reproduction that results from the reformatting process (and sometimes more). And our identifiers can easily exceed 32 characters. Therefore, in our final published guideline, we departed from the EBU specification and recommended placing one or more tagged identifiers in the 256-character bext *Description* element. This conflicts with the EBU specification, which defines *Description* as an "ASCII string…containing a free description of the sequence. To help applications which only display a short description, it is recommended that a résumé of the description is contained in the first 64 characters, and the last 192 characters are use for details."

When promoting a guideline or standard, one of the issues to address concerns the ease with which the user community can comply: are there tools for the job? After publishing our guideline for metadata in the EBU bext chunk and the RIFF/WAVE INFO chunk, we asked our expert consultants from AudioVisual Preservation Solutions to produce an edit-and-embedding tool. The resulting software package is named *BWF MetaEdit* and it has been pilot-tested by three federal agencies. We plan to place it on the SourceForge website during the summer of 2010 as an open-source offering to all interested archives.

As we drafted our guideline, we found that we were not alone in facing header anemia. When the Still Image Working Group developed their initial guideline for embedding metadata in image files, they started with the TIFF header and found that the options for identifier embedding were limited and there was no good way to identify certain details, e.g., an image's color space (except in rather general terms) or a scanning device's color profile. The shortfalls encountered while developing WAVE and TIFF guidelines have motivated both Working Groups to explore additional approaches to embedding metadata. For example, the Audio-Visual Working Group plans to revisit the two underused WAVE-related specifications from EBU: aXML and iXML.

One option for the Still Image Working Group is the useful ANSI/NISO Z39.87 standard, *Data Dictionary – Technical Metadata for Digital Still Images*. The development of Z39.87 by NISO was itself motivated in part by a perception of TIFF header anemia. The Z39.87 standard offers several dozen data elements that document technical features at the file level. The XML manifestation for this data set is called *NISO Metadata for Images in XML* (MIX). Since most archives implement this data set using MIX as an extension schema of METS, however, most expressions of Z39.87 metadata are managed in package-level metadata sets and are not embedded in image files directly, our quest at the moment.

The Still Image Working Group is exploring XMP (eXtensible Metadata Platform), an open specification for embedded, file-level metadata from Adobe. XMP is supported by the widespread availability of tools from Adobe and others, most of which permit both the creation of the data and its automated migration within the family of common images formats, e.g., TIFF, PDF, GIF, PNG, SVG, JPEG, and JPEG 2000. Easy metadata migration would be very helpful in a reformatting program that creates masters in one format and derivative images in another. The group noted that many professional photographers make use of the combined metadata specifications of XMP and IPTC, the data set standardized by the International Press Telecommunication Council. Incidentally, IPTC picture data includes elements for multiple identifiers.

Moving image formatting

Recommendations and guidelines should follow and reflect experience, and the Working Group has been tracking the progress being made by the three federal agencies that have begun to digitally reformat analog, standard definition videotapes. Our interest, however, is by no means limited to standard definition video. All of our participating agencies look forward to digitally reformatting high definition video and motion picture film in a few years' time and we seek an extensible approach to formatting.

To date, the Library of Congress has done the most digital video reformatting while the National Archives and Records Administration and the Smithsonian Institution are starting to carry out projects of their own. All three agencies have purchased SAMMA devices, a product of the Front Porch Digital company. The Library is using SAMMA's best-known implementation in a workflow that produces a stream of video-frame images, each encoded in lossless JPEG 2000. This picture data, together with soundtrack, timecode, closed captioning, and so on, is wrapped in the Material eXchange Format (MXF) file format. Files in this format serve as archival masters for preservation in the moving image collections at the Packard Campus for Audio-Visual Conservation, Culpeper, Virginia.

JPEG 2000 is a standard from the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). MXF is a standard of the Society of Motion Picture and Television Engineers, and some refer to it as a *container* or a *wrapper*. The growing body of experience with MXF-wrapped JPEG 2000 files means that this is an important target format for the Working Group to consider. At the same time, we are tracking other video reformatting efforts, including a trio of activities that entail the capture and storage of uncompressed video streams. One of these is at Stanford University, another at Rutgers, and a third at the BBC. The BBC approach is of special interest because it also employs the MXF container format.

What are the Working Group's impressions thus far? First, we see merit in exploring an approach based in the MXF standard, which is seeing increasing adoption in the professional broadcast industry, and in JPEG 2000 picture encoding, which is also seeing increasing adoption in various moving image sectors, e.g., as part of the digital cinema specification. Nevertheless, we want to keep an eye on uncompressed picture encoding as well, especially in examples like the one from the BBC, with wrapping in MXF.

Second, we are aware that MXF and JPEG 2000 are broad-spectrum standards that feature many options for packaging, metadata, and encoding. The successful implementation of an approach that uses these standards—and/or uncompressed video encoding, for that matter—will be enhanced if we users define a set of constraints. Well-defined constraints will support the development of tools to validate files and encourage multiple vendors to provide conforming equipment. A documented set of constraints increases the level of standardization applied to digital content, which in turn increases interoperability, content exchange, and long-term, preservation-oriented data management.

For users of the MXF standard, formal constraint statements are called *Application Specifications*. These can be compared to JPEG 2000 *profiles* or to the *profiles* and *levels* that characterize MPEG video content. The incubation of MXF Application Specifications is the special province of the Advanced Media Workflow Association, an organization that provides a meeting ground for professional moving-image users and vendors. Our Working Group plans to work with the AMWA to define one or more preservation-oriented Application Specifications.

The development of an application specification for moving image preservation will benefit from the involvement of archives beyond our federal agencies. For this reason, the Working Group is planning a technical meeting on digital-video-reformatting target formats to coincide with the joint conference of the International Association of Sound and Audiovisual Archives (IASA) and the Association of Moving Image Archivists (AMIA) in Philadelphia in November 2010. Technically oriented persons from interested organizations who wish to attend should contact the writer of this article for more information.

Conclusion

The examination—one might even say unpacking—of formatting elements for sound recordings and moving image content highlights the many, complex facets that must be considered. The Working Group's investigation points to the high value of documents like profiles and application specifications that supplement published standards for important formats. Such documents provide a detailed record of what is being produced, thus supporting the interoperability of content between organizations and over time. Finally, as the snapshots in this article show, preservation *practices* will be built upon many *standards* from many sources. Both federal agencies Working Groups hope to offer guidelines for good practices that reference well-chosen standards. | IP |

doi: 10.3789/isqv22n2.2010.07

CARL FLEISCHHAUER <cfle@loc.gov> is a Program Officer for the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress.

RELEVANT LINKS

Advanced Media Workflow Association Application Specifications
www.amwa.tv/projects/application_specifications.shtml

aXML, EBU 3285 Supplement 5
tech.ebu.ch/docs/tech/tech3285s5.pdf

Broadcast WAVE Audio File Format, version 1, 2001
tech.ebu.ch/publications/tech3285

Broadcast Wave Metadata
www.digitizationguidelines.gov/audio-visual/documents/wave_metadata.html

Data Dictionary – Technical Metadata for Digital Still Images, ANSI/NISO Z39.87
www.niso.org/standards/z39-87-2006/

Digital Cinema System Specification
www.dcimovies.com/

Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format, ISO 12234-2:2001
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29377

Embedding Metadata in Digital Audio Files: Introductory Discussion for the Federal Agencies Guideline
www.digitizationguidelines.gov/audio-visual/documents/Embed_Intro_090915.pdf

EXIF Specifications
www.exif.org/specifications.html

eXtensible Metadata Platform (XMP)
www.adobe.com/devnet/xmp/

Federal Agencies Digitization Guidelines Initiative
www.digitizationguidelines.gov

File-based Production: Making It Work in Practice, BBC Research White Paper, WHP 155, September 2007
www.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP155.pdf

Graphic technology – Prepress digital data exchange – Tag image file format for image technology (TIFF/IT), ISO 12639:2004
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34342

Guidelines on the Production and Preservation of Digital Audio Objects, 2nd ed., IASA TC-04
www.iasa-web.org/audio-preservation-tc04

Information technology – JPEG 2000 image coding system: Core coding system, ISO/IEC 15444-1:2004
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37674

International Press Telecommunications Council (IPTC) Photo Metadata
www.iptc.org/IPTC4XMP/

iXML
www.gallery.co.uk/ixml/compatible.html.

Linear Pulse Code Modulated Audio (LPCM)
www.digitalpreservation.gov/formats/fdd/fdd000011.shtml

Material Exchange Format (MXF)
www.digitalpreservation.gov/formats/fdd/fdd000013.shtml

MBWF / RF64: An Extended File Format for Audio, EBU-TECH-3306-2007
tech.ebu.ch/docs/tech/tech3306-2009.pdf

METS Schema & Documentation
www.loc.gov/standards/mets/mets-schemadocs.html

MIX: NISO Metadata for Images in XML (MIX)
www.loc.gov/standards/mix

MPEG-2, Video profiles and levels
en.wikipedia.org/wiki/MPEG-2#Video_profiles_and_levels

Multimedia Data Standards Update, April 15, 1994
www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/Docs/RIFFNEW.pdf

Multimedia Programming Interface and Data Specifications 1.0, IBM Corporation and Microsoft Corporation, August 1991
www.tactilemedia.com/info/MCI_Control_Info.html

PhotoMetadata.org
www.photometadata.org

RUCore: Rutgers Community Repository, Recommended minimum standards for preservation sampling of moving image objects, April 6, 2007
rucore.libraries.rutgers.edu/collab/ref/dos_avwg_video_obj_standard.pdf

Safeguarding the Audio Heritage: Ethics, Principles, and Preservation Strategy, Version 3, IASA TC-03
www.iasa-web.org/content/safeguarding-audio-heritage-ethics-principles-preservation-tc03

SAMMA products, Front Porch Digital
www.fpdigital.com/Products/Migration/Default.aspx?mrsc=MigOverview

Self-documentation as a Sustainability Factor
www.digitalpreservation.gov/formats/sustain/sustain.shtml#self

Sound Directions project
www.dlib.indiana.edu/projects/sounddirections/papersPresent/sd_bp_07.pdf

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

Still Image Working Group, Content Categories and Digitization Objectives
www.digitizationguidelines.gov/stillimages/documents/ccdo-subcat-T1.html

Still Image Working Group, Embedded Metadata subgroup charter
www.digitizationguidelines.gov/stillimages/sub-embedded.html

Tagged Image File Format (TIFF) 6.0 specification
partners.adobe.com/public/developer/tiff/

WAVEFORMATEXTENSIBLE
www.microsoft.com/whdc/device/audio/multichaud.mspx

What is a Format
www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Digital Preservation in Capable Hands:

Taking Control of Risk Assessment at the National Library of New Zealand

KEVIN DE VORSEY AND PETER MCKINNEY

New Zealand's digital documentary heritage is encoded according to a diverse array of file formats. Identification and characterization of the formats is a constant challenge. This challenge makes it difficult to establish an accurate risk view of the content to mitigate format obsoleteness.

The National Digital Heritage Archive (NDHA) of the National Library of New Zealand Te Puna Mātauranga o Aotearoa has concluded that the measurement of conformance of files to a format standard for such risk analysis is at best insufficient and at worst harmful. For the digital documentary heritage of New Zealand, the ideal is the measurement of individual file profiles against application specifications. This gives a meaningful and actionable risk view of our content.

With no limitations or control over the format of the content that is collected and preserved, the Library has issues to resolve before the long-term preservation of digital collections can be assured. There are many significant obstacles that make the term "permanent access" an almost meaningless catchphrase when applied to such a collection of digital content made up of disparate file formats. Solving these and other problems is the responsibility of the National Digital Heritage Archive (NDHA) and a significant step has been taken through the development of the Rosetta preservation repository system in conjunction with Ex Libris Group.

While the life-span of content stored on physical materials such as paper, glass, wood, and stone can be accurately predicted based on hundreds of years of experience, backed by scientific research into material composition and the effects of environmental conditions like temperature and humidity, the best that the preservation community can do with digital material is to make educated guesses based on a few decades of mostly anecdotal experience. The concept of information encoded according to a file format has only been in existence since about the 1950s and therefore the field of digital preservation must be considered as being still in its infancy. Happily, significant advances have occurred in the area of data storage and management that permit cultural heritage institutions to manage enormous digital collections of permanently valuable material in online (or nearly online) repositories of spinning disks and/or robotic tape libraries. Through the use of checksums to detect format rot or corruption, virus scanning to protect against malicious code, robust network and physical security, and comprehensive disaster planning, it is not too far-fetched to believe that it is now possible to guarantee bitstream preservation—which is to say, preserving deposited files perfectly in their original form. We view this as "passive preservation" that is

CONTINUED »

foundational to digital preservation. Unfortunately, while the perfect preservation of a human-readable format such as a paper manuscript is usually synonymous with access to its content, bit-preservation of electronic formats is not. The inevitable obsoleteness of the hardware and software components necessary to interpret and render files in a usable form makes it necessary to complement perfect but passive preservation with some form of active, managed preservation. (We are painfully aware that we do not discuss in more detail our use of the word "render." It is a loaded term with many levels of interpretation. We are currently defining this internally as it is critical to our risk analysis. Space deters us from exploring it further in this paper.) This demands an accurate risk view of the repository. This risk view is the mechanism that offers enough warning to the NDHA in order that action can be taken to allow continued access to the material.

The problem with format specification adherence as an indicator of risk

From our reading, the primary methods currently being suggested for predicting this type of risk involve the comparison of files to a format specification, which in turn is graded against agreed-upon sustainability criteria. There exist many misunderstandings around format sustainability that have contributed to the idea that there are "archival" or "preservation" formats. The NDHA is uncomfortable with the concept of inherently preservation-worthy formats. It is our belief that while sustainability factors may prove useful for a forensic understanding of formats in the future and in interpreting files that are discovered after a period of benign neglect, there are other more practicable methods for identifying risk that are better suited for supporting active preservation in a repository.

Along with bit preservation, accurate identification of a file's encoding (its format) is foundational to preservation. In response, the preservation community has developed utilities that identify files by format and "validate" them; that is, measure their compliance with the format specification with which they have been associated. It is certainly useful to possess the information that a string of bits is an audio file that is encoded according to the Broadcast Wave EBU Specification and that they failed JHOVE's measures of well-formedness or validity. But, it is more important to know that the bits were written according to a profile that has been associated with a particular legacy application, and that this application is known to encode in a non-standard way due to an aspect of the specification that was originally open to interpretation. The NDHA has encountered this phenomenon with a number of formats including Rich Text Format, Tagged Image File Format, and Broadcast Wave.

The content the NDHA preserves

The National Library can, and does accept all formats. It collects content, not "perfect" formats. All materials selected as Library collection items are ingested into the preservation repository essentially as is. The current policy of the NDHA is not to transform content into preferred formats on ingest, but this may be considered after additional research is conducted. In order to actively preserve this content, we must first understand exactly what form it is in. Every file must therefore be identified by its format and where possible, a picture created of its characteristics. Once this is done, different management views can be taken. The range of formats is very wide. We have Sibelius music composition files, web harvests in ARC format, Wordstar, MacWrite, TIFF, JPEG, GIF, text, mp3, flac, wave, broadcast wave, and a whole host of format unknowns.

Through the use of checksums to detect format rot or corruption, virus scanning to protect against malicious code, robust network and physical security, and comprehensive disaster planning, it is not too far-fetched to believe that it is now possible to guarantee bitstream preservation—which is to say, preserving deposited files perfectly in their original form.

The range of formats the National Library accepts is very wide:

- wave, broadcast wave
- Wordstar
- TIFF, JPEG, GIF
- ARC format
- Sibelius music composition
- MacWrite
- text, mp3, flac
- unidentified

Our experience with New Zealand's documentary heritage is that files contain multifarious properties. These are based on the world of possibilities that the format standard describes, but can also include non-standard properties. The range of possibilities and relationships between them is such that it is quite meaningless to purely measure a file's adherence to the format standard.

We can take PDF files as an instructive example. Adobe has clear standards for versions of PDF. However, the wide range of applications that can create PDFs do not always stick to this standard. Indeed, non-adhering PDFs can be made by Adobe's own suite of applications. What does it mean if a PDF is invalid because it does not have tags for the images as the standard requires? Should we base risk on this non-compliance?

In addition, consider this: it is not a bold statement to suggest that the majority of the world-wide Web is written in non-conforming code. Would this content not be at risk if it was written in perfect code? Is the conformance of the code really the biggest risk facing this material?

NDHA Risk Analysis

Within the NDHA, we base risk on practical capabilities; risk analysis through tracking format standards is too abstract for us. There has been a degree of literature about risk analysis of collections. Our understanding of the body of work is that many have coalesced around the utilization of what are described as "sustainability factors." Depending on the source, these number from seven to fourteen. The original study by Arms and Fleischhauer identifies seven factors. These were put in place to assess the sustainability of formats for preserving content. Further work at the Dutch National Library moved this work into the area of risk assessment for their own very specific circumstances (regular access is not offered by the institution to the materials this was applied to). In essence across all the literature on this, the criteria remain essentially the same but are given different situational groupings and nomenclature. They include factors such as the level of documentation for a format, a format's backwards compatibility, and its complexity.

We do not believe these factors belong in the area of risk assessment of collections. Within our repository, we do note sustainability factors against formats and applications. But these are not used to determine any view of risk to content due to format obsoleteness. (For an excellent discussion on the terms obsolete, obsolescence, and obsolescent, see Pearson & Webb 2008.) We will use those factors to offer decision-making information when selecting formats to migrate content into (i.e., dealing with the risk). However, it is unlikely that they will be key elements of decision-making as the most important input will be the new format's ability to render and our level of comfort with that rendering.

We can take PDF files as an instructive example. Adobe has clear standards for versions of PDF. However, the wide range of applications that can create PDFs do not always stick to this standard. Indeed, non-adhering PDFs can be made by Adobe's own suite of applications.

The risk analysis method the National Library employs measures each ingested file against a format and application relationship. Simply, our view of risk is that if the National Library cannot render it, it is at risk. We use a format, application, and risk library within Rosetta to identify risks. As the first stage of risk analysis, the format and application libraries use two levels of relationship—an association and an activation. A format can be associated with an application. For example, we can link PSD files with Adobe Photoshop CS3. However, even with this association, any PSD files we receive will still be classed as at risk. The second level of linkage is an activation of the association. An activation is the institution (in this case, NLNZ) declaring that they have the application within their own environment. ("Environment" here means "within the institution." The activated application could be embedded within the preservation system, deployed as part of the institution's IT infrastructure, or even held on a stand-alone PC within a relevant business unit. Critically though, the application is under the control of the institution.) Once this activation is made, then PSD files in the repository are no longer viewed as being at risk: the National Library can render the content.

What these relationships between the format and application libraries offer is a description of the universe of rendering possibilities and the ability to define a world view upon which an institution's risk is determined. This is the most basic level of risk analysis employed in Rosetta.

A more detailed interpretation is the next layer of risk analysis. This layer looks at the characteristics of the files themselves. What we do is understand the capabilities of the applications we have and determine if there are certain

CONTINUED »

What does it mean if we have 10,000 images in a format that a risk matrix based on sustainability factors tells us are at risk because documentation on the format is incomplete? If we can happily render these files, this analysis is unhelpful.

properties that will cause them to "reject" a file that is otherwise in the correct format. These properties are then noted in the risk library. It is important to understand that we do not maintain a registry of all characteristics that are possible within a format; we only note exceptions that break the format/application relationship.

For example, the Library receives a number of MP3 files each week. We know currently that we have the rendering capability for MP3s encoded with LAME and Fraunhofer methods. However, we cannot reliably render MP3s that are created using the Xing encoding method. If a Xing MP3 is deposited, the relationship at the format and application level determines that the file is not at risk, because at this gross level, we are happy with the file. However, the extracted metadata (from the NLNZ metadata extraction utility) contains the troublesome encoding. This does not stop the file from being passed to the permanent repository, but the next day, when the risk report is re-run it identifies this particular file as matching the risk profile and reports on it as being at risk.

Conclusion

Where does this leave us? If we rely on identification and characterization tools that measure a file's risk through adherence to a standard, we are basing our risk on what to us, is relatively meaningless information. What does it mean that our TIFF is invalid? What does it mean if we have 10,000 images in a format that a risk matrix based on sustainability factors tells us are at risk because documentation on the format is incomplete? If we can happily render these files, this analysis is unhelpful. The most worrying end of this particular road is that it could very well guide us to a course of action when none is actually required.

The risk we have been discussing is the risk of what the community terms "digital obsolescence." It is our view that risk is situational; it is not a statement of fact. At risk is not an inherent state of files and formats, it is an institution's view of its content determined by the policies, guidelines, and drivers it has at any one point in time. We have included no discussion on our "control" of the applications used to render files. Tracking the contract dates and review dates for all applications is our method of analyzing "obsolescence" (the march to being obsolete). This will give is adequate time to plan for action that is truly required.

The National Library of New Zealand, by basing its risk routines on institutional rendering capability, creates a view of its repository that gives accurate and meaningful information on what can and cannot be rendered. To us, this is the essence of obsoleteness. | IP | doi: 10.3789/isqv22n2.2010.06

KEVIN DE VORSEY <Kevin.DeVorsey@natlib.govt.nz> is an Electronic Formats Specialist at the National Records and Archives Administration, but at time of writing was the Digital Preservation Analyst in the NDHA. Peter McKinney (Peter.McKinney@natlib.govt.nz) is NDHA Policy Analyst at the National Library of New Zealand Te Puna Mātauranga o Aotearoa.

RELEVANT LINKS

Arms, Caroline and Carl Fleischhauer. Digital Formats: Factors for Sustainability, Functionality, and Quality. IS&T Archiving 2005 Conference, Washington, D.C.
memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf

European Broadcast Union Specification of the Broadcast Wave Format
tech.ebu.ch/docs/tech/tech3285.pdf

Lawrence, Gregory W., et al. Risk Management of Digital Information: A File Format Investigation. Council on Library and Information Resource, June 2000.
www.clir.org/pubs/reports/pub93/contents.html

National Digital Heritage Archive (NDHA)
www.natlib.govt.nz/about-us/current-initiatives/ndha

NLNZ Metadata Extraction Tool
meta-extractor.sourceforge.net/

Rog, Judith and Caroline van Wijk. Evaluating File Formats for Long-term Preservation. National Library of the Netherlands, 2008.
www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf

Rosetta system, Ex Libris
www.exlibrisgroup.com/category/RosettaOverview

Strodl, Stephan, et al. How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In: International Conference on Digital Libraries, Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries, Vancouver, BC, Canada: 2007.
www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf

Pearson, David and Colin Webb. Defining File Format Obsolescence: A Risky Journey. The International Journal of Digital Curation, 3(1), 2008.
www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf

OP [OPINION]

A judgment formed about something;
a personal view, attitude, or appraisal

Mary Molinaro

MARY MOLINARO

How Do You Know What You Don't Know? Digital Preservation Education

Two Scenarios: Scanning Projects Gone Bad

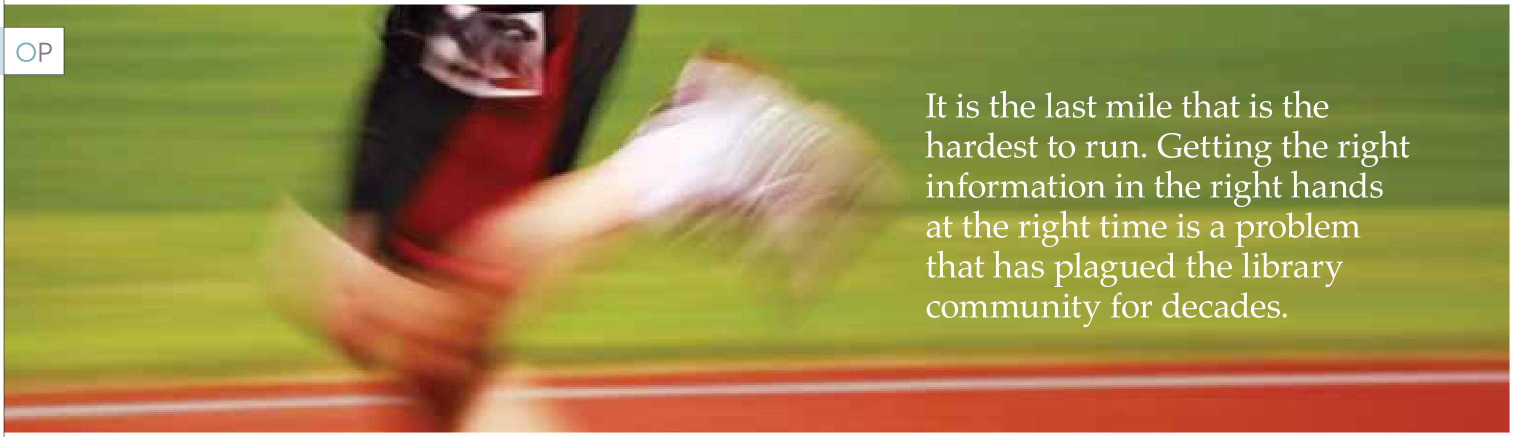
Imagine this scenario: a curator for a local history museum is approached by the museum director to scan some of the photo collections and make an online exhibit. The museum has a web page and the director suggests the photos be put on that page somewhere. The museum has a flatbed scanner and the curator goes to work scanning. The collection of 100 photographs takes quite a bit of time to scan, but within a couple of weeks the images are scanned. The curator has some experience with webpages and places low-resolution copies of the images on a webpage linked from the museum's main page. The JPEG copies are on the hard drive of the computer attached to the scanner and are numbered sequentially starting with IMG001.jpg. The curator realizes that the images should be preserved and so copies the files onto gold CDs so they will be safe. In reality, the curator clearly does not understand archival file formats, the intricacies of content management systems, issues with file naming conventions, or that CDs are an unstable and impermanent storage media.

In another scenario the Press Association of a medium-sized state is interested in having the state's newspapers made available online. They are aware that there are large runs of microfilm in the state historical society. In addition, there are large numbers of other state documents that would also be useful. The historical society recently purchased a state of the art microfilm scanner and has tested it enough to know that the scanner is very fast and very good. When approached by the Press Association about scanning the film, they estimate how long it will take them to scan all 200,000 reels of film and with the new scanner realize that it will not take very long at all. They agree to do the job for $200,000. Once they start the project they quickly realize that the files that are created are quite large; so large they can't afford the storage to store the TIFF images. They also realize that they have not planned for a way to present the pages to users other than as a series of JPEG images. There is also no preservation plan for the images. Rather than go back to the Press Association to re-scope the project, the director of the society decides to do the best they can now and make improvements later—after all it is digital access and it is better than nothing.

In reality, a poorly conceived plan is not better than nothing. Spending limited resources on projects that will have little hope of being sustainable is a tremendous waste that serves no one well. Unfortunately, scenarios similar to these are playing out all across the country. Yes, there are many well thought-out projects with preservation plans in place, but in so many organizations a little knowledge about scanning and webpages can be a dangerous thing. Every institution with responsibility for the stewardship of materials in digital form has some interest in long-term digital preservation. How are the staff members in organizations across the country expected to have the knowledge and skills to ensure that their projects and programs are well conceived, feasible, and have a solid sustainability plan? In short, how does the staff know what they do not know?

CONTINUED »

How are staff members in organizations across the country expected to have the knowledge and skills to ensure that their projects and programs are well conceived, feasible, and have a solid sustainability plan?



It is the last mile that is the hardest to run. Getting the right information in the right hands at the right time is a problem that has plagued the library community for decades.

Overview of efforts

Since 2000 there have been at least two surveys of preservation readiness in cultural heritage organizations in the U.S., both published in 2005. In 2003 Cornell University began surveying 100 institutions participating in its Digital Preservation Management Workshops and in April 2005 the Northeast Document Conservation Center (NEDCC) surveyed 169 cultural heritage institutions on a wide variety of topics related to digitization and digital preservation. Those surveys revealed that barely a third of the respondents had policies in place for the management and preservation of digital content. (As a point of fact, since 2007 the Inter-University Consortium for Political and Social Research (ICPSR) has partnered with Cornell to offer the Digital Preservation workshops and tutorial with support coming from the National Endowment for the Humanities since 2008. They continue to collect data from participants on digital readiness.)

Since those surveys were completed there have been numerous educational opportunities that include significant information in scanning standards and preservation planning that enable practitioners to gain experience in scoping digital projects. Notably the School for Scanning, Digital Directions, and the Persistence of Memory workshops offered by the Northeast Document Conservation Center (NEDCC) and the Digital Preservation Management Workshop formerly at Cornell and now sponsored by the ICPSR.

There are also numerous conferences such as the International Conference on Preservation of Digital Objects (iPRES) or the International Digital Curation Conference (IDCC) that are held for practitioners and center on the topic of digital preservation.

There are some excellent programs being offered by the iSchools around the country to educate new librarians and archivists to teach the skills needed as they move into professions steeped in digital expectations. The programs at the University of Arizona, the University of Michigan, and the University of North Carolina at Chapel Hill are examples of those who are turning out grads who both understand the issues and who will be prepared to lead the way as they move into positions across the country. These new professionals will be highly desirable for the skills they bring to the table in terms of digital acumen.

Additionally, the University of Arizona's graduate certificate program in Digital Information Management (DigIn) and the University of North Carolina's DigCCur program that has the tag line, "Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum" are offering both education and the development of communities of practice for working practitioners. The School of Information at the University of Michigan has a program to create internship opportunities in digital preservation, administration, and curation. All three programs receive support from the Institute for Museum and Library Services (IMLS).

Standards, Tools, and Projects

Over the last decade there has also been movement in terms of standards and best practices for digitization and sustainability. There are a number of resources available to provide guidance for those undertaking digital projects. In fact, when looking for guidance there are many "imaging guidelines" available from a wide variety of organizations. But most are highly technical and many are out of date. It is understandable if people actually undertaking digital projects set these aside in favor of the manual that came with the scanner or the advice of a well meaning colleague. One must know that their current practice is lacking to even look for improvements.

Additionally, a great deal of effort has gone into the development of tools to assess the strengths and weaknesses of existing repositories of digital data. Of note are the Trustworthy Repositories Audit and Certification (TRAC) criteria and checklist that was developed by RLG and the National Archives and Records Administration (NARA) and the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA). These tools can be used for repository planning as well as assessment.

The National Digital Information Infrastructure and Preservation Program (NDIIPP) has as its mission "to develop a strategy to collect, preserve and make available significant digital content, especially information that is created in digital form only, for current and future generations." With this mission comes the realization that this will require effort at the local level so that material is available to preserve. NDIIPP is funded by Congress and is leveraging the weight of the Library of Congress to begin to mobilize at a local level. Partners across the country have been involved in many worthy and important initiatives including MetaArchive, the Internet Archive, LOCKSS, and Portico. With the Digital Preservation website, NDIIPP presents an excellent set of resources for librarians, archivists, and the public.

The Last Mile

With all of these resources, why is it that the librarians in the local libraries are still making grave and costly errors in building sustainable collections when asked to do a "scanning project"? Why are the tools and standards being largely ignored at the most basic level in so many institutions across the country? How is it with the wealth of information and training available to the library community that it seems so elusive to so many people?

I believe the analogy often given to describe so many types of projects is very true in this case: it is the last mile that is the hardest to run. Getting the right information in the right hands at the right time is a problem that has plagued the library community for decades. When adding in the incredible pace of change in the digital environment, limited resources for training and travel, and work days that are already overburdened, it is not surprising that at the local level people forge ahead on projects blissfully unaware of standards and best practices.

In this area of rapid change it is those who are already heavily involved in the development of the tools and services that are best able to leverage their use. Unfortunately there are still vast numbers of people and project managers who have no idea of where they should even start.

Moving Forward: A call to action

Something has got to give. If we have any hope to preserve the digital record of our lives and collections there must be a coordinated effort that takes advantage of the years of work that has been put into the development of the practices that will provide the best shot at sustainability. People at the local level must be encouraged and supported to represent their collections and communities in a digital form that has a very good chance to persist over time. We must leverage the expertise that exists and make it easy for people at the local level to know what to do.

To this end the Library of Congress, through the NDIIPP program, is taking a leadership role once again. Initial steps have been taken to establish a broad-based education program to reach practitioners across the country through a program dubbed Digital Preservation Outreach and Education (DPOE). This program is in the planning stage, but the idea of taking training and education for digital preservation into the heart of the country will make all the difference in empowering the front lines in the fight for sustainability of our digital heritage. Updates will be available on the NDIIPP website as the planning unfolds.

It is important for those who are knowledgeable to participate in ways that will make a real difference. Partnerships and collaborations will fit hand in glove with education programs offered at the local level. In the digital preservation community we have talked around these issues for many years. Increased visibility at a local level supported by national organization will finally make it possible for all of the talk to become reality. | OP | doi: 10.3789/isqv22n2.2010.08

MARY MOLINARO <molinaro@uky.edu> is Director, Preservation and Digital Programs, in the University of Kentucky Libraries.

RELEVANT LINKS

Clareson, T. "NEDCC Survey and Colloquium Explore Digital Preservation Policies and Practices." RLG DigiNews Feb 15, 2006.
worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file1339.html#article1.

DigCCurr, University of North Carolina at Chapel Hill,
www.ils.unc.edu/digccurr/

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

Digital Preservation Workshop, Inter-University Consortium for Political and Social Research (ICPSR)
www.icpsr.umich.edu/dpm/

International Conference on Preservation of Digital Objects (iPRES) 2010
www.ifs.tuwien.ac.at/dp/ipres2010/

International Digital Curation Conference (IDCC), 2010 (6th)
www.dcc.ac.uk/events/conferences/6th-international-digital-curation-conference

Kenney, A. and E. Buckley. "Developing Digital Preservation Programs: the Cornell Survey of Institutional Readiness, 2003-2005." RLG DigiNews, Aug 15, 2005.
worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file1088.html#article0

National Digital Information Infrastructure and Preservation Program (NDIIPP), Library of Congress.
www.digitalpreservation.gov/library/

National Digital Information Infrastructure and Preservation Program (NDIIPP) Initiatives
www.digitalpreservation.gov/library/initiatives.html

Northeast Document Conservation Center (NEDCC) workshops
www.nedcc.org/education/workshops.introduction.php

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

University of Arizona, Graduate Certificate Program Digital Information Management
digin.arizona.edu/index.html

SP [SPOTLIGHT]

MARILYN REDMAN AND LAURA MCCARTHY

National Archives and Records Administration: The Nation's Recordkeeper

Marilyn Redman
Management and
Program Analyst

Laura McCarthy
Senior Policy Analyst

What better NISO member to spotlight for this issue's theme of preservation than long-time NISO voting member, the U.S. National Archives and Records Administration (NARA). Marilyn Redman <marilyn.redman@nara.gov>, Management and Program Analyst, and Laura McCarthy <laura.mccarthy@nara.gov>, Senior Policy Analyst, responded to questions from the ISQ editor about NARA and their involvement with standards and preservation.

Q For readers who aren't familiar with NARA, can you briefly explain what the agency does?

The National Archives and Records Administration (NARA) is the nation's recordkeeper. We preserve, safeguard, and make available the records of our Government, ensuring that the people can discover, use, and learn from this documentary heritage. We establish policies and procedures for the preservation and management of U.S. Government records; manage the Presidential Libraries system; and publish Federal laws and regulations, as well as Presidential and other public documents.

Q How do people who don't work for the government benefit from NARA's preservation activities?

In a democracy, the records of the Government belong to its citizens, and providing access to them is a vital service. Working with Federal agencies as our partners, the Archivist and NARA staff identify records to be retained for posterity. NARA then gathers, stores, processes, and preserves the records. Our holdings can only be made available to current and future generations if we invest in the archival preservation and processing of records in our custody.

Q Tell us about NARA's Electronic Records Management (ERM) and Electronic Records Archives Program (ERA).

Electronic records management [ERM] guidance at NARA is created in two units of Modern Records Programs: the scheduling and appraisal division and the ERM policy team. These two units produce front-end records management guidance for use by Federal agencies ranging from advice on scheduling requirements to advice on format selection and technology-specific management guidance. Guidance is frequently produced in consultation with archival custodial units as well as selected Federal agencies and the Chief Information Officers Council.

The Electronic Records Archives (ERA) Program is NARA's strategic response to the challenges posed by electronic records, and it is providing the foundation for an e-government approach to the management of all types of Federal records. ERA supports NARA's mission for oversight of the management of records by all agencies of the U.S. Government. It also enables NARA to preserve and provide access to increasing volumes of historically valuable electronic records, in ever more complex formats. ERA is being implemented as a set of federated systems. To date, three instances of ERA have been deployed. The first provides online tools for agencies to request and receive authorization from NARA for disposition of Federal records. That instance is being expanded and enhanced for the preservation of permanent electronic records. The second instance was designed to handle the over 72 terabytes of electronic presidential records from the George W. Bush White House, which were transferred to NARA on January 20, 2009. The third version is customized for electronic records of the Congress. Work is currently underway to develop additional capabilities for preservation and access that will be available in all instances of ERA.

Q How has NARA incorporated standards into its services and which standards (NISO or others) are most important to NARA?

NARA has incorporated ISO 15489 (*Information and documentation – Records management – Part 1: General*) as an underlying tenet in the recent update of our regulations for Federal records management (36 CFR Chapter 12 Subpart B). NARA guidance and training emphasizes the principles contained in ISO 15489-1. We believe that this standard is useful to all records managers.

NARA leverages its involvement in the PDF/A standard to inform our transfer instructions for Permanent Records in PDF. Participating in PDF/A provides NARA with a comprehensive technical understanding of the PDF file format. This helps us to develop transfer instructions that restrict use of PDF features that could complicate the long term preservation of information maintained as PDF.

Q What benefits has NARA gained from utilizing standards and incorporating them into its services?

The National Archives and Records Administration (NARA) is involved in many standards activities focusing on ensuring long-term usability, authenticity, and preservation of records over time. NARA's standards participation touches on a wide range of issues relating to both electronic and physical records in the following areas:

- Records management – usability and authenticity
- Digital repositories and digital preservation – management, maintenance, and preservation of electronic or digital records
- Interoperability and IT environments – data exchange, management, and storage
- Stability and storage of physical records – stability, permanence, and storage of paper, photographic materials, and electronic storage media

Additionally, basing citations in our regulatory environment on international standards, and incorporating such citations into our training and guidance, demonstrates to our stakeholders a long-standing commitment to international collaboration. We continue to believe that approaches based on voluntary consensus standards offer the best opportunity for leveraging experiences in differing juridical environments which provides alternative approaches to those we might otherwise develop.

Q What standards development has NARA been actively involved in and what benefits do you gain from involvement in standards development?

NARA has been involved in the development of a number of international standards. As a member of ISO TC46/SC 11 (Information and documentation/Archive & records management), NARA was involved in the development of ISO 15489, the original core international records management standard. Following on that work, NARA has also actively participated in the development of ISO standards and technical reports relating to metadata, business process analysis, records management compliance, archival description, and preservation. NARA is now participating in the development of an ISO management system family of standards relating to records management. This management system of standards will stand beside ISO 9000 and ISO 14000 as fundamental management systems for operating in the global economy. NARA is also a participant on the Joint Working Group responsible for ISO 19005-1:2005 (*Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*).

NARA was heavily involved in the development of the *Open Archival Information System Reference Model* (ISO 14721:2003) and is contributing to the development of the follow-on standard, *Digital Repository Audit and Certification*, being developed by the Consultative Committee on Space Data Systems for submission to the ISO under TC20/SC13 (Aircraft and space vehicles/Space data and information transfer systems). NARA is also collaborating with the National Institute of Standards and Technology (NIST) in the

CONTINUED »

The burgeoning use of social media tools represents a significant challenge to records management principles and techniques. We are evaluating the recordkeeping aspects of specific social media technologies, on a case-by-case basis, and would welcome an international perspective in evaluating and addressing these challenges.

new initiative to develop a standard for a digital preservation interoperability framework under ISO/IEC JTC1 Study Group on Digital Content Management and Protection. NARA contributed to the development of the PREMIS metadata model for digital preservation. NARA has also been working for several years with several other archives, libraries, and cultural institutions in the effort to create a Universal Digital Format Repository (see article on page 26), which will provide a shared source of basic data on the great variety of digital formats that need to be preserved.

In addition to ISO standards participation, NARA has also been involved with standards development work of the International Council on Archives, the Object Management Group, the IEEE, and the World Wide Web Consortium.

Q What problem areas have you encountered that would benefit from further standards or best practices development?

The burgeoning use of social media tools represents a significant challenge to records management principles and techniques. We are evaluating the recordkeeping aspects of specific social media technologies, on a case-by-case basis, and would welcome an international perspective in evaluating and addressing these challenges.

A second issue for NARA is the difficulty of keeping track of the many standards activities that occur in the various ISO technical committees. As a result, we are sometimes surprised when draft standards appear in near final form that can have a significant impact on our mission. This problem also exists for standards activities that occur in organizations outside the ISO framework. RSS feeds for mission related subjects might help alleviate the laborious and sometimes unsuccessful efforts to scan the horizon for relevant ongoing work.

Q What else would you like NISO ISQ readers to know about NARA?

Promoting and ensuring effective records and information management across the Federal Government is the foundation on which the long-term success of NARA's mission depends. We carry out this foundational work by ensuring that:

- Federal agencies can economically and effectively create and manage records necessary to meet business needs,
- records are kept long enough to protect rights and assure accountability, and
- records of archival value are preserved and made available for future generations. | SP |

doi: 10.3789/isqv22n2.2010.09

RELEVANT LINKS

NARA website
www.archives.gov

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

IEEE Standards Association
standards.ieee.org/

International Council on Archives standards
www.ica.org/en/standards

Object Management Group specifications
www.omg.org/gettingstarted/overview.htm

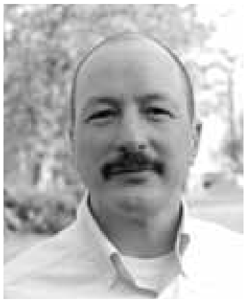
Permanent Records in PDF
www.archives.gov/records-mgmt/initiatives/pdf-records.html

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

World Wide Web Consortium
www.w3.org

Jim LeBlanc



Adam Chandler
E-Resources & Database
Management Research
Librarian, Cornell University

JIM LEBLANC

Measuring the Quality of OpenURLs: An Interview with Adam Chandler

NISO's Business Information Topic Committee approved in December 2009 the establishment of a new working group called IOTA—Improving OpenURL Through Analytics. Chaired by Adam Chandler, E-Resources & Database Management Research Librarian in Central Library Operations at Cornell University, the working group will build on work previously conducted by Adam at Cornell. Jim LeBlanc, Director of Delivery & Metadata Management Services and Adam's colleague at Cornell, talked to him about the work he had already done and the follow-up project at NISO.

Q Let's start with something simple, Adam. What are OpenURLs?

Back in the 1990s, the only way to link from an article citation to a full text document was through something called bilateral linking. Each vendor needed to pre-compute and maintain all the links between their site's content and every other vendor site they linked out to. Then Herbert Van de Sompel and his colleagues at Ghent University came along and figured out a way to pass metadata to software that knows something about a library's collection, a method to exchange information to help a patron answer the question: does the library have access to this resource—print or electronic—and if so where is it? They essentially moved the job of maintaining the links to a brand new node in the supply chain, one optimized for the task: the "link resolver." Then they proposed a standard for the syntax of this "OpenURL" that would allow for predictable transfer of the resource's metadata.

The development of OpenURLs was hugely successful, because it addressed what was known as the "appropriate copy problem," a term that refers to the inadequacy of standard URLs to lead a user from the citation of an article to the most suitable full-text copy of that article. Commercial link resolver software was developed in the early 2000s to take an incoming OpenURL and: (1) determine if the library has a subscription to the journal in question, and (2) if so, present a new URL to the library patron that will connect him or her to full text—or to the library catalog or an interlibrary loan request form, if full text is not available. In 2004, the original OpenURL specification was generalized into a formal standard, ANSI/NISO Z39-88:2004, *The OpenURL Framework for Context-Sensitive Services.*

Q What's your specific interest in OpenURLs and quality metrics?

OpenURL was a genuine breakthrough and innovation for libraries. In 2009, Cornell patrons alone clicked on about half a million OpenURL citation links. In a talk last year, Herbert mentioned that a conservative estimate is that over a billion OpenURL requests are made by library patrons every year. The access these links provide can be very satisfying for library patrons, but bad links can be extraordinarily frustrating. Many vendors offer OpenURL links on their sites, but after the links go out to library link resolvers, the vendors have no idea what happens. They get no systematic feedback and don't know if library patrons are able to successfully access resources from their links. The aim of my project is to devise a method to provide feedback to vendors regarding the quality of the metadata content they're sending out, because the reality is OpenURLs don't work 100% of the time. Some OpenURL providers are better at supplying complete and accurate data than others. Nobody knows how often patrons are successful when they click on an OpenURL.

Q Where are you now in your research?

I've been gathering up usage log files from different link resolvers from three different institutions and three vendors. I have complete data for 2009 from Cornell, the Georgia Institute of Technology, and Kansas State University, plus sample data from EBSCO, Serials Solutions, and Thomson Reuters—a total of over 4,475,000 OpenURLs. I've written a program that parses each OpenURL, counts the elements that

CONTINUED »

are most likely to be needed for successful linkage (title, ISSN, author, date, and so forth), and indicates whether these elements are present or absent in the OpenURLs. Within each field of the OpenURL, I look for other things, such as whether dates have been entered in the correct form. The results are loaded into a database from which anyone can request reports.

Q Can content providers request reports on the quality of their own OpenURL data?

The web reporting system is currently organized by the institution or vendor who supplied the link resolver log file and date, but it is possible to generate an offline report for a vendor. For example, a year and a half ago Eric Rebillard, Professor of Classics and History at Cornell and editor of the bibliographic database *L'Année philologique*, was getting a number of complaints about failed OpenURL links. David Ruddy, Cornell University Library's Director of Electronic Publishing, and I worked with Eric to obtain a planning grant from Mellon to improve links from *L'Année*. The primary focus of the grant was experimental work on something called canonical citation linking. A secondary focus of the proposal was to develop an automated method for evaluating OpenURL quality. Eric is currently working with his programmers to fix the problems we identified when we ran the 900,000 plus OpenURLs through the parser. I recently ran a sample of OpenURLs for another vendor, the American Institute of Physics. I look forward to working with more vendors, as more of them find out about the NISO initiative.

Q So it's necessary to keep the vendor-supplied data separate from other data in the database?

I believe it is. The data from the 900,000 citations in *L'Année*, for example, would distort the results from other queries on the database. The point of the current system is to be able to pull data from library link resolvers for a specified time period (quarterly), because we want to monitor changes in quality over time. As vendors are sensitized to the issues and can see how their own OpenURLs compare in quality to those of their peers, they will, I hope, allocate resources to fix the problems that are uncovered. We will write a report on the efficacy of this model after two years and make a recommendation on its continuation. If vendors fix problems, we'll consider the work a success. If they ignore them, well, I might conclude that there is an inherent flaw in the OpenURL linking model that probably won't be fixed.

Q What's next?

The OpenURL standard has been around for ten years now, but this is the first attempt to create a feedback loop to help improve the quality of the data passed along in OpenURLs. A related issue is how to improve the proprietary and nonstandard inbound linking from the link resolver to the full-text content provider sites. I've been working on this problem for a while, thanks to our collaboration with Professor Rebillard, but the NISO initiative is helping me bring in other collaborators and solicit more interest in the issue. We have a great group of experts on board for the NISO project; members include Susan Marcin from Columbia University, Oliver Pesch from EBSCO, Ellen Rotenberg from Thomson Reuters, Elizabeth Winter from Georgia Tech, and Rafal Kasprowski from Rice. The existing OpenURL standard was developed under the aegis of NISO, so it makes sense to develop the quality metrics within the structure of NISO. Working through NISO will also keep the process transparent and impartial. We're also working closely with the joint NISO/UKSG KBART (Knowledge Base And Related Tools) Working Group that is developing recommended practices to improve OpenURL knowledge bases.

Q What can people do who are interested in the project?

We'd like to get more log files of OpenURL linking from those managing a link resolver, whether library or vendor. Anyone who has data they would be willing to share can contact me <email: alc28@cornell.edu>. An interest group e-mail list is available for anyone who wants to follow the activities of the working group. | NR | doi: 10.3789/isqv22n2.2010.10

JIM LEBLANC <jdl8@cornell.edu> is Director of Delivery & Metadata Management Services, Cornell University Library.

RELEVANT LINKS

Canonical Citation Linking and OpenURL
cwkb.org

Improving OpenURL Quality Through Analytics (IOTA) Working Group
www.niso.org/workrooms/openurlquality

OpenURL Quality Metrics Database
openurlquality.niso.org

OpenURL Quality Metrics Interest Group E-mail List
openurlqualityinfo-subscribe@list.niso.org

CR [CONFERENCE REPORT]

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

CONTINUED »

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.

| CR | doi: 10.3789/isqv22n2.2010.11

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.

RELEVANT LINKS

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

Melissa Goertzen

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

Valerie Ryder

[CONFERENCE REPORT]

VALERIE RYDER

Electronic Resources & Libraries 2010 Conference

The 5th Electronic Resources & Libraries (ER&L) Conference was held February 1–3, 2010, in Austin, TX. This conference is planned by academic librarians to discuss issues concerning electronic resources and to share best practices.

The 350 attendees from 40 states and six countries were primarily from the academic library community (77%) with 16% representing the information industry and the remainder from libraries in government agencies, institutions, and corporations. Registration is capped to maintain the collegial atmosphere of the conference and to facilitate networking and discussion. Vendors from the information content and services industry provide some funding for the conference to defray expenses and keep the registration fee as low as possible, compared with other conferences. The only sales opportunity is the two-hour Sponsors' Reception, hosted by the vendors, on Monday evening.

Prevalent themes of this year's conference were:

» ways of dealing with library budget reductions (a relatively new experience for academic librarians),

» using electronic resources usage statistics to evaluate the collections and determine cancellations,

» creation and maintenance of digital collections with local content (such as historical documents, maps, artifacts, and art objects),

» standards for data and system interfaces, and

» technology trends such as electronic books (e-books), mobile access devices, web-scale discovery, and federated search.

The overwhelmingly favorite topic on attendees' minds from the Thought Cloud produced on the ER&L website prior to the conference was Usage Metrics. Workflow and E-books tied for second place in the myriad of issues that attendees wanted to discuss.

Pre-conference seminars on January 31 covered how to successfully market electronic resources, techniques for processing, storing, and viewing usage data and explored current dilemmas in information ethics.

The keynote speaker, University of Texas School of Information professor **Lance Hayden**, set the tone for the conference by challenging attendees' thinking about security and privacy as related to digital information in the wild world of the Internet.

The 45 sessions presented over two and a half days were organized into ten tracks:

1. Electronic Resource Management (ERM) Systems
2. Managing Electronic Resources
3. Standards
4. E-books
5. Statistics & Assessment
6. ER Delivery & Promotion
7. Scholarly Communication
8. Collaboration
9. Emerging & Future Technologies
10. Collection Development

Most of the presentations were case studies discussing how these issues are addressed at the presenters' institutions and offering best practices and lessons learned.

Academic librarians are struggling with the significant reductions in their subscriptions budgets that have occurred each year since 2007. Many libraries subscribe to full-text journal databases that duplicate coverage with their print journal collections so they are paying for the same journal title multiple times, sometimes with different time period coverage and embargos. Librarians are struggling with how to identify and eliminate the overlaps to reduce their subscription spending without reducing their content scope. **Diane Carroll** of Washington State University, **Tim Jewell** of the University of Washington, and **Nina Bakkalbasi** of Yale University Library

each presented their data-intensive work processes for assessing the journal collections at their universities. **Gayle Baker** and **Ken Wise** of the University of Tennessee presented how they calculate Return on Investment (ROI) for their journal collections as part of justifying their budget levels and proving value to their institutions.

Many of the libraries are embracing the e-book as another delivery mechanism for content whether the media is a web-based e-book, a digital book delivered through the Internet, a digital book delivered to a mobile device (e-book reader or other device), or a web-based Major Reference Work. Libraries are subscribing to e-book collections as well as individual book titles. They are challenged by the selection and evaluation process concerning which media and platform to choose as well as accommodating these media into their ordering and cataloging processes. **Lee Hisle** of Connecticut College, **Ellen Safley** of the University of Texas at Dallas, and **Nancy Gibbs** of Duke University shared their experiences with patron-driven selection of e-books where students and faculty members determine which e-books are added to the library collection through their usage of specific e-books from collections made available in the library's online public access catalog. **Dani Roach** and **Carolyn DeLuca** of the University of St. Thomas delivered humorous but realistic insights on the quandary that many e-books have the characteristics of serials and databases as well as print books. Their interactive dialogue with attendees illustrated how librarians are rethinking and revising their technical work processes to handle e-books as hybrids. During session question and answer periods as well as networking conversations, some librarians expressed their preferences for buying individual e-books by title rather than collections of e-books. They commented that they lost the ability to select e-journals by title when they converted to the e-journal bundles and they do not want to lose that selectivity when they move to e-books from print books. The lament of "life was easier in the print media for books" was expressed by more than a few librarians.

Many of the libraries are embracing the e-book as another delivery mechanism for content whether the media is a web-based e-book, a digital book delivered through the Internet, a digital book delivered to a mobile device (e-book reader or other device), or a web-based Major Reference Work.

A common theme throughout the conference was that librarians want systems to handle all their electronic resources from selection and procurement to user access and delivery. They want new systems, such as electronic resource management systems (ERMs), to integrate well with their existing systems, such as their integrated library system (ILS) or online public access catalog (OPAC). In many cases, the academic librarians have tested the ERMs from many vendors but found them lacking desired features or incompatible with their existing work processes, so they have developed their own ERM system or particular modules. Those libraries that have implemented vendor ERM systems have had to modify their work processes to conform to the system's requirements. Excellent case studies were presented by **Abigail Bordeaux** of

CONTINUED »

Web-scale discovery engines have to normalize metadata harvested from many sources in order for search to work well. Vendor web-scale discovery engines are still being improved and may not have normalized metadata for all sources that a library wants to search.

Harvard University as well as **Benjamin Heet** and **Robin Malott** of University of Notre Dame.

The standards scene for ERM systems is still in flux. **Tim Jewell** of the University of Washington gave a good review of the NISO ERM Working Group's fast track efforts to perform a "gap analysis" of the remaining management and data standards issues. Original standards such as Electronic Resource Management Initiative (ERMI) were not standards but really "pre-standards." Vendors found it hard to develop systems using these standards. Following the analysis, the working group will make recommendations regarding the future of the ERMI data dictionary and identify gaps in interoperability and best practices to inform future work.

Another issue that was discussed frequently was the proliferation of problems with the quality of data passed from content providers to link resolvers, even those adhering to the OpenURL standard. Cornell University's **Adam Chandler** presented results of a study of several link resolvers and the data being passed from content providers. The success of an OpenURL link resolver finding the right article depends upon which data elements are passed from the content provider. A two-year NISO project was approved in December 2009 to investigate the feasibility of creating industry-wide transparent and scalable metrics for evaluating and comparing the quality of OpenURL implementations across content providers.

Interest continues in web-scale discovery engines as an improvement over federated search. **George Boston** of Western Michigan University explored the advantages and challenges of each approach to providing the user with a simple, easy, and fast search solution that unifies all of the resources in a library. Web-scale discovery engines have to normalize metadata harvested from many sources in order for search to work well. Vendor web-scale discovery engines are still being improved and may not have normalized metadata for all sources that a library wants to search. This is similar to the lag in developing the search maps for each source that federated search engines needed to crawl when they were first marketed.

Jamene Brooks-Kieffer of Kansas State University Libraries challenged attendees to envision the future with new services and evolving standards as the journal article becomes the primary entity of scholarship. Technology solutions will expedite the discovery to delivery process for users but mask the librarian's role in ensuring seamless integration and provide business model challenges to information industry players who have a stake in the current workflow.

Closing session panelists provided the transition to next year's ER&L by discussing tools and technologies for the future. **Andrew Nagy** of Serials Solutions focused on Software as a Service (SaaS) and cloud computing, library resources discovery services, and next generation catalogs. **Ross Singer** of Talis illustrated the concept of linked data by showing how difficult it is for users to see how the library's collection of data is connected to the vast world of external data.

Once again the ER&L Conference provided a venue to share knowledge and experiences with electronic resources, learn of new developments and potential solutions, and debate challenging ideas in an open dialogue between the library community and information industry partners.

VALERIE RYDER <vryder@wolper.com> is Director of Information Strategy at Wolper Subscription Services in Easton, PA.

RELEVANT LINKS

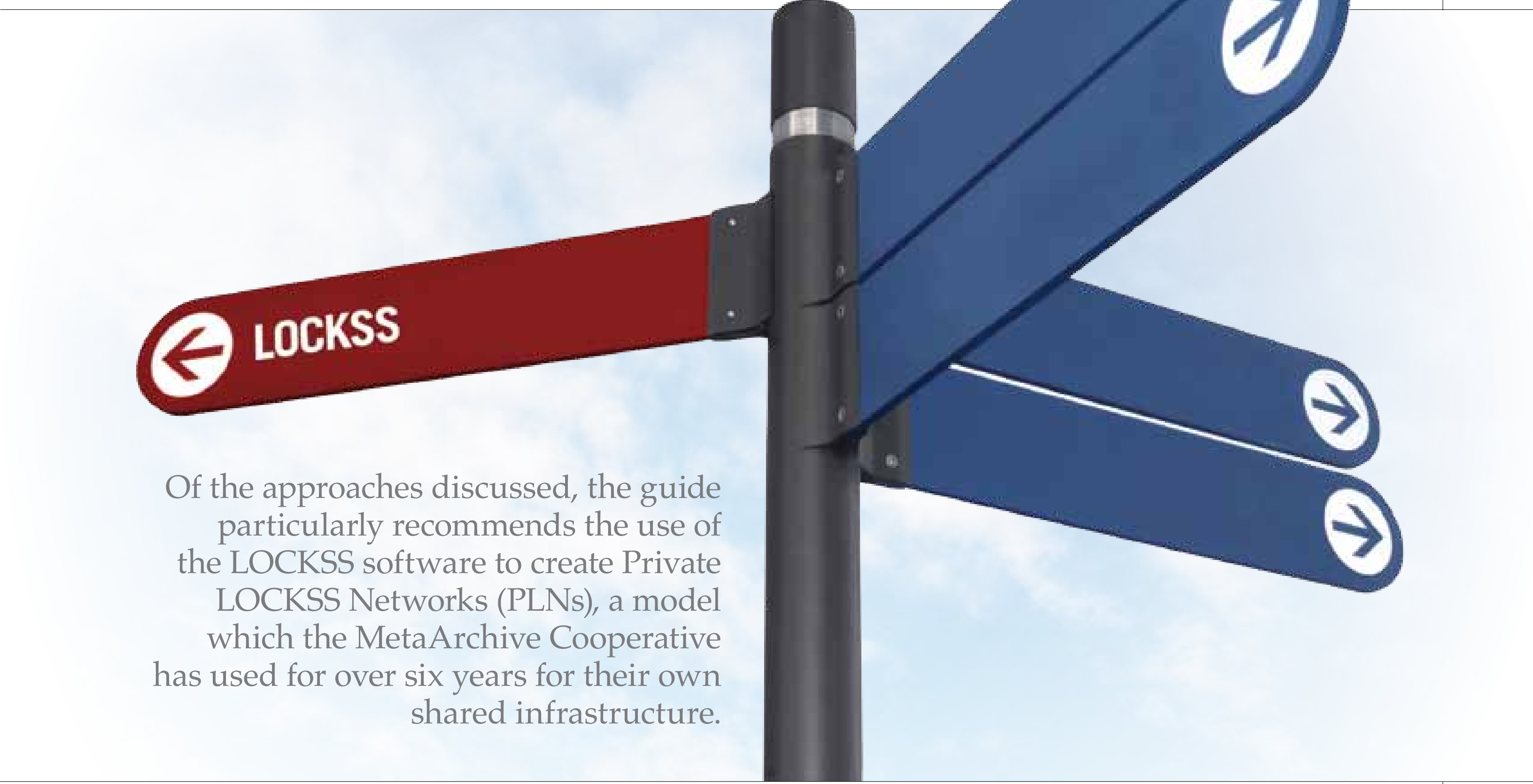
Electronic Resources & Libraries 2010 website
www.electroniclibrarian.org/erlwiki/ER%26L

Electronic Resources & Libraries 2010 presentation materials
www.electroniclibrarian.org/erlwiki/Program#Presentations_Listed_by_Track

NISO ERM Data Standards & Best Practices Review
www.niso.org/workrooms/ermreview

NISO Working Group on OpenURL Quality Metrics
www.niso.org/workrooms/openurlquality

NW [NOTEWORTHY]



Of the approaches discussed, the guide particularly recommends the use of the LOCKSS software to create Private LOCKSS Networks (PLNs), a model which the MetaArchive Cooperative has used for over six years for their own shared infrastructure.

Distributed Digital Preservation Guide from the MetaArchive Cooperative

Touted as being the first in a series of preservation publications, *A Guide to Distributed Digital Preservation*, produced by the MetaArchive Cooperative, contains a collection of articles by Cooperative members on the still-emerging field of using digital replication and distribution for preservation.

Of the approaches discussed, the guide particularly recommends the use of the LOCKSS software to create Private LOCKSS Networks (PLNs), a model which the MetaArchive Cooperative has used for over six years for their own shared infrastructure. The PLN model is used as the underlying framework and case example for the Guide's discussions of architecture; technical considerations; organizational considerations; content selection, preparation, and management; content ingest, monitoring, and recovery; cache and network administration; and copyright practices—although much of the content is considered extensible to other distributed digital preservation solutions as well.

In providing the guide, the authors hope to forestall a trend towards outsourcing of digital collection management. "The central assertion of the MetaArchive Cooperative… is that cultural memory organizations can and should take responsibility for managing their digital collections, and that such institutions can realize many advantages in collaborative long term preservation and access strategies." A distributed strategy with institutional collaboration and investment supported by a robust technical infrastructure is the proposed alternative. ■

Download the guide at: www.metaarchive.org/GDDP

Planets Survey Gauges Organizational Readiness for Digital Preservation

The Preservation and Long-Term Access Though Networked Services (Planets) Project has issued a white paper summarizing a market survey of over two hundred organizations, mainly European archives and libraries, investigating their digital preservation activities and needs.

Findings of the survey included:

- The volume of digital content that organizations expect to archive will increase 25-fold over the next ten years. By 2019, 70% of survey respondents expect to hold over 100TB of content.
- Within a decade, over 70% will need to preserve video, audio, databases, websites, and e-mail in addition to the current needs for documents and images.
- The three most important capabilities of a digital preservation system were: maintaining the authenticity, reliability, and integrity of records; checking that records have not been damaged; and planning the preservation of content to deal with technical obsolescence.
- Open-source and proprietary software are used equally by respondents, and often combined in the same solution.
- Respondents were much less interested in emulation than in migration as a preservation solution for technology obsolescence.
- Only 27% think that they have complete control over the file formats that they will accept and store in their digital archives.
- Compliance with metadata standards is regarded as fairly important, but there is less agreement on which standards. Dublin Core was the most popular (used by 51% of respondents), followed by MARC (31%) and ISAD(G) (28%).
- Organizations are only starting to commit to funding digital preservation; just 47% have allocated a budget to it.
- There is evidence that digital preservation is emerging as a profession in its own right; where previously the work was carried out by IT and preservation or curation staff, now it is starting to be carried out by specialists.

The white paper concludes with a summary of needed next steps including the importance of having a digital preservation policy and the need for more and better tools to automate the preservation process. ■

The Planets white paper and a separate survey analysis report are available from: www.planets-project.eu/publications/

Report on Sustainable Economics for Long Term Preservation

For its final report, The Blue Ribbon Task Force on Sustainable Digital Preservation and Access has taken an economic perspective on all the resources—human, technical, and financial—that are needed to ensure that digital assets will be available for future use. *Sustainable Economics for a Digital Planet* identifies three requirements:

1 articulate a compelling value proposition;

2 provide clear incentives to preserve in the public interest;

3 define roles and responsibilities among stakeholders to ensure an ongoing and efficient flow of resources to preservation throughout the digital lifecycle.

National and international agencies, funders, and sponsors of data creation, stakeholder organizations, and individuals are all called upon to take particular actions to ensure preservation and access.

Four domains—scholarly discourse, research data, commercially owned cultural content, and collectively created web content—were analyzed for their sustainability risks and domain-specific recommendations were made. For example, in the domain of scholarly discourse, the Task Force recommended that "publishers reserving the right to preserve should partner with third-party archives or libraries to ensure long-term preservation." For collectively produced web content, the Task Force suggests that "creators, contributors, and host sites could lower barriers to third-party archiving by using a default license to grant nonexclusive rights for archiving."

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access was created in late 2007 with funding from the National Science Foundation and The Andrew W. Mellon Foundation, in partnership with the Library of Congress, the Joint Information Systems Committee of the United Kingdom, the Council on Library and Information Resources, and the National Archives and Records Administration. ■

All of the Task Force's publications, including *Sustainable Economics for a Digital Planet*, are available for free download from: brtf.sdsc.edu/publications.html

Keeping Research Data Safe

Building on the Phase 1 report's cost model for the long-term preservation of research data, the Keeping Research Data Safe (KRDS) project has just published their Phase 2 report, which reports the results of testing and validating that cost model. Survey cost data was received on 13 collections and the cost information from four organizations—Archeology Data Service, National Digital Archive of Datasets, UK Data Archive, and University of Oxford—were analyzed in depth and presented in case studies. A benefits framework was also developed and illustrated with case studies from the National Crystallography Service at Southampton University and the UK Data Archive at the University of Essex.

Among the report's conclusions were:

- The costs of acquisition/ingest and access are far greater than the costs of the archiving activities. Thus, it is likely that the largest potential cost benefits will come from the development of tools that support the ingest and access activities.
- Once core fixed costs are in place (largely staff resources), increasing levels of economies of scale can be demonstrated as content is added.
- The documentation of the dataset can be as beneficial to archive as the data itself. (In one example, the documentation was downloaded 10 times more often than the actual dataset.)
- The OAIS reference model fits better with preservation services focused on data archives and institutional repositories. It is less ideal for focusing on "near-term preservation and curation work from a researcher perspective."
- The benefits taxonomy has great potential for further development and implementation. [Ed. Note: While the study focused on datasets, the benefits taxonomy could be easily transferable to other types of preservation collections and activities.]

The KRDS2 study was funded by JISC with support from OCLC Research and the UK Data Archive. ■

The full report can be downloaded from the KRDS2 webpage: www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx

JHOVE2 Beta Released

A new beta version of JHOVE2, the open-source Java framework for format-aware characterization of digital objects, has been released to the public. Characterization not only provides information about an object but can also function as a surrogate, which is especially useful in preservation environments. JHOVE2 uses the processes of identification, validation, feature extraction, and assessment to derive the characterization.

JHOVE2 is being designed as a next-generation improvement to the original JHOVE software, based on over four years of extensive use. While the original JHOVE assumed that one object was equivalent to one file and one format, JHOVE2 supports a single object having multiple files and formats. The redesigned architecture also configures the modules so they can be iteratively applied to each object. Among the other improvements are a plug-in interface, de-coupling the identification and validation, performance enhancements, and better error reporting.

This beta release has been provided to give interested users an early look at the new JHOVE2 architecture and APIs. While the processing modules are fully functional, there is limited format support at this time. Additional format modules will be added as they are completed.

The JHOVE2 project is a collaborative undertaking of the California Digital Library, Portico, and Stanford University, with funding from the Library of Congress as part of its National Digital Information Infrastructure Preservation Program. ■

For further information and to download the beta release, visit the JHOVE2 website: https://confluence.ucop.edu/display/JHOVE2Info/

Codecs Primer for Archives

AudioVisual Preservation Solutions has published *A Primer on Codecs for Moving Image and Sound Archives: 10 Recommendations for Codec Selection and Management* by Chris Lacinak. In addition to providing introductory material on what encoding and compression are and how they work, the paper emphasizes that the choice of these schemes can impact the ability to preserve the digital object.

Since moving images and sound generally require a codec for the decoding process, the selection process is critical, whether dealing with born digital content, reformatting older content, or converting analog materials. Ten recommended approaches are explained: adoption, disclosure, transparency, external dependencies, documentation and metadata, pre-planning, maintenance, obsolescence monitoring, maintenance of the original, and avoidance of unnecessary transcoding or re-encoding.

There is no single "right" codec; each archive needs to make the decision as part of an overall preservation strategy. ■

The Codecs Primer can be downloaded from: www.avpreserve.com/wp-content/uploads/2010/04/AVPS_Codec_Primer.pdf

SS [STATE OF THE STANDARDS: *May 1, 2010*]

In Development or Revision

Listed below are the NISO Working Groups that are currently developing new or revised standards, recommended practices, or reports. Refer to the NISO website (www.niso.org/workrooms/) and *Newsline* (www.niso.org/publications/newsline/) for updates on the Working Group activities.

WORKING GROUP	STATUS
Cost of Resource Exchange (CORE) Co-chairs: Ed Riding, Ted Koppel	**Z39.95-200x, Cost of Resource Exchange (CORE) Protocol** Draft Standard for Trial Use (DSFTU)
DAISY/NISO Standard Advisory Committee Chair: George Kerscher	**Z39.86-201x, Specifications for the Digital Talking Book** Standard revision in development.
E-Journal Presentation & Identification Co-chairs: Cindy Hepfer, TBA	Working group roster formation underway.
ERM Data Standards & Best Practices Review Co-chairs: Ivy Anderson, Tim Jewell	Technical Report in development.
Institutional Identifiers (I^2) Co-chairs: Grace Agnew, Oliver Pesch	**Z39.94-201x, Institutional Identifiers** Standard in development.
IOTA: Improving OpenURLs Through Analytics (formerly OpenURL Quality Metrics) Chair: Adam Chandler	Technical Report in development.
Knowledge Base and Related Tools (KBART) Phase II *Joint project with UKSG* Co-chairs: Andreas Biedenbach, Sarah Pearson	**NISO RP-9-2010, KBART: Knowledge Bases and Related Tools** Issued January 2010. Phase II Recommended Practice development work now underway.
Physical Delivery of Library Materials Co-chairs: Valerie Horton, Diana Sachs-Silveira	Recommended Practice in development.
RFID for Library Applications Revision Co-chairs: Vinod Chachra, Paul Sevcik	**RP-6-201x, RFID in U.S. Libraries** Revision in development.
Single Sign-on (SSO) Authentication Co-chairs: Steve Carmody, Harry Kaplanian	Recommended Practice in development.
Standardized Markup for Journal Articles Co-chairs: Jeff Beck, B. Tommie Usdin	**Z39.96-201x, Standardized Markup for Journal Articles** Standard in development.
Supplemental Journal Article Materials Co-chairs Business Working Group: Linda Beebe, Marie McVeigh Co-chairs, Technical Working Group: Dave Martinsen, Alexander (Sasha) Schwarzman	Working group roster formation underway.

2010 Library Assessment Conference

Building Effective, Sustainable, Practical Assessment

Baltimore, Maryland | October 25-27, 2010

Register Now!

Keynote Themes	Speakers
Library Service Quality	Fred Heath, University of Texas
Performance Measures and Balanced Scorecard	Joe Matthews, San Jose State University
Assessment of Library Spaces	Danuta Nitecki, Drexel University
Information Literacy	Megan Oakleaf, Syracuse University
Value and Impact	Stephen Town, The University of York, UK

The conference goal is to support and nurture the library assessment community through a mix of invited speakers, contributed papers and posters, workshops, and engaging discussion.

Join our ongoing discussion on library assessment issues, visit the library assessment blog or subscribe to arl-assess@arl.org.

www.libraryassessment.org

University of Virginia Library

Association of Research Libraries

W University Libraries
University of Washington