ARTICLE EXCERPTED FROM:

ISQ INFORMATION STANDARDS QUARTERLY

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

SPECIAL ISSUE: DIGITAL PRESERVATION

DIGITAL PRESERVATION METADATA STANDARDS

TRUSTWORTHY DIGITAL REPOSITORIES

UNIFIED DIGITAL FORMATS REGISTRY

AUDIO-VISUAL DIGITIZATION GUIDELINES

DIGITAL PRESERVATION EDUCATION

NISO
How the information world CONNECTS

DIGITAL PRESERVATION

METADATA STANDARDS

A publication of the National Information Standards Organization (NISO)

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

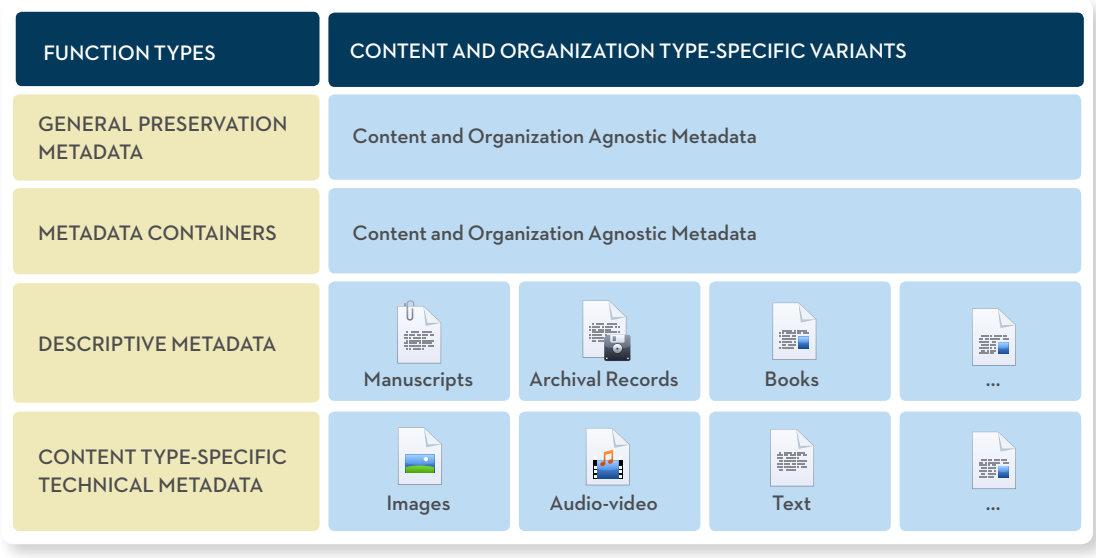


FIGURE 1: The Space of Digital Preservation Metadata Efforts

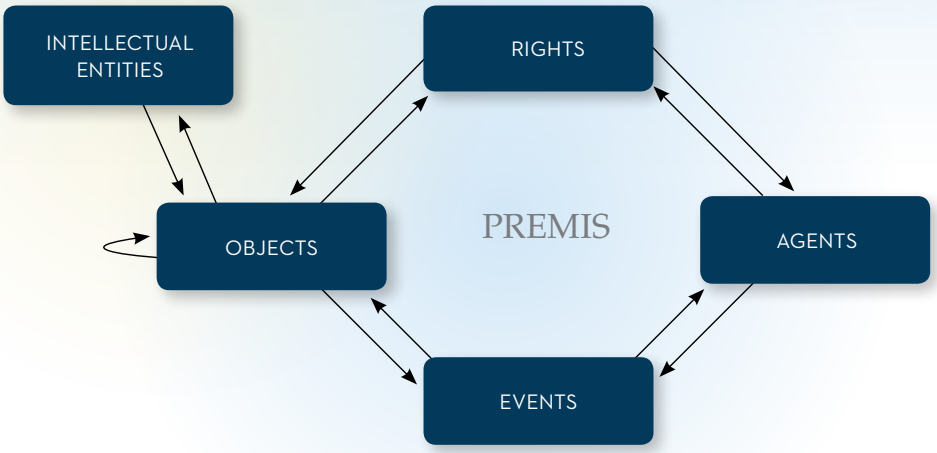


FIGURE 2: The PREMIS Data Model

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.*

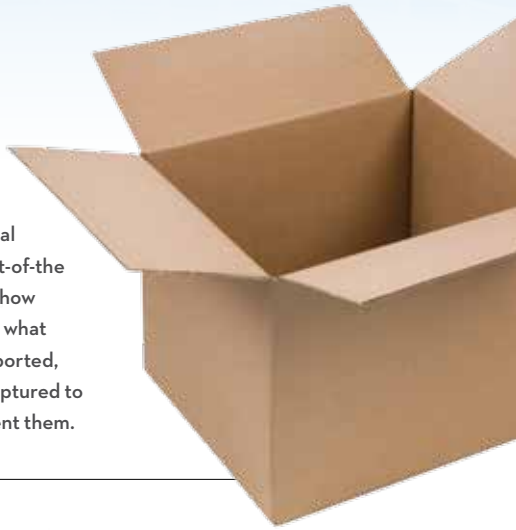
Given the wide range of institutional contexts, PREMIS cannot be an out-of-the box solution. Users have to decide how to model their specific application, what business functions need to be supported, which semantic units need to be captured to support them, and how to implement them. In addition, they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

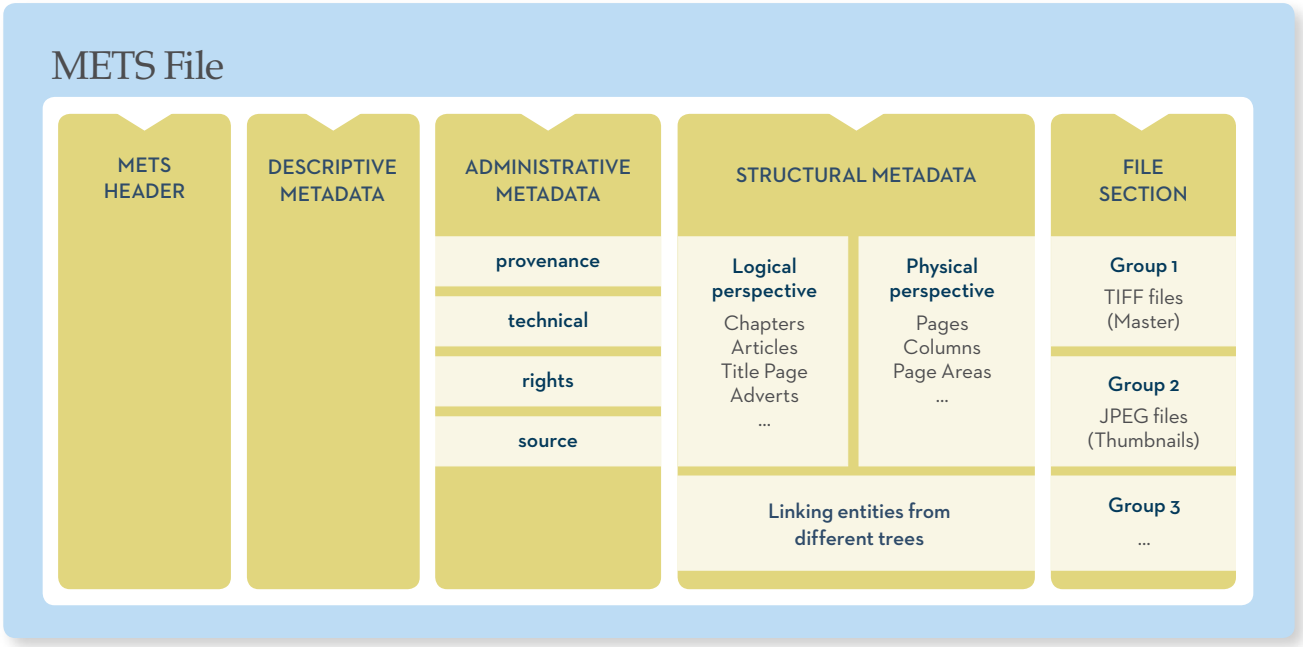


FIGURE 4: The METS Architecture

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.

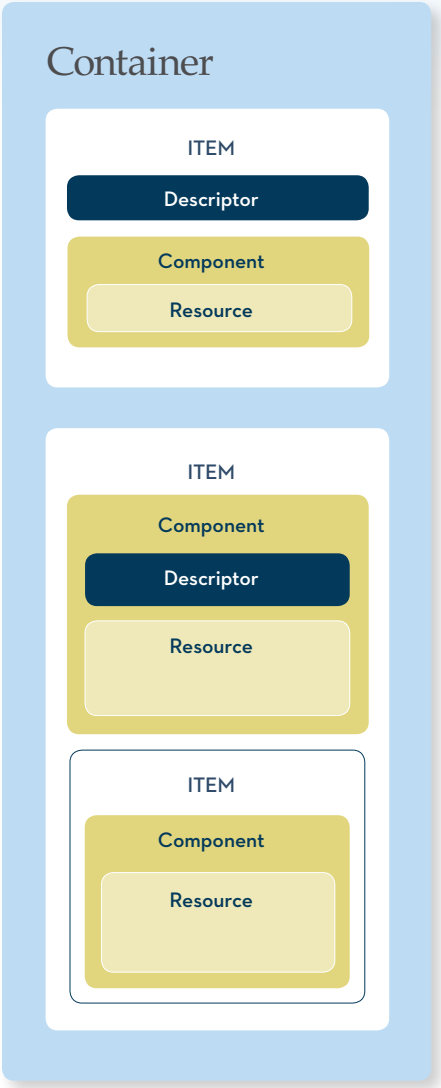


FIGURE 5: The MPEG-21 DIDL Architecture

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 »

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.

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.

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