TSW 02-05

October 2002

METS: Metadata Encoding and Transmission Standard

Richard Gartner

Pearson New Media Librarian

Oxford University Library Services

Executive Summary

Recent years have seen the publication of several standards designed to encode metadata for objects held within digital library collections: what has been lacking is an overall *framework* within which these schemes can be integrated. The **Metadata Encoding and Transmission Standard (METS)** attempts to fill this role: it is an newly emergent standard designed to encode metadata for electronic texts, still images, digitized video, sound files and other digital materials within electronic library collections. In doing so, it attempts to address the lack of standardization in digital library metadata practices which is currently inhibiting the growth of coherent digital collections.

Written in XML schema, METS offers a coherent overall structure for encoding all relevant types of metadata (descriptive, administrative, and structural) used to describe digital library objects. It also allows for the encoding of specific behaviours necessary for the rendering of these objects. METS allows this metadata to be either embedded directly within its own structure, or held in external files and referenced from within it. An extensive system of internal identifiers is used to link all facets of the digital object's components together. By virtue of being written in XML, METS is platform- and software independent, robust and readily interchangeable with other schemes.

METS is the first widely-accepted standard designed specifically for digital library metadata: other XML applications have been used for this purpose before, but all were initially designed to fulfill other functions and so do not readily adapt to this one. To supplement the overall framework provided by METS, its editorial board recommend various “extension schemas”, which are used to extend it to incorporate specific types of metadata in a standardized way.

As an XML schema application, METS files can be created, administered and delivered by any number of XML-compliant software tools, most of which are free of charge. A number of stylesheets have already been written to render and deliver METS files in specific applications, and others produced to convert legacy XML metadata to METS. In addition, a number of software applications which will facilitate the administration of METS files are under development.

Despite its recent provenance, METS has already been adopted by a number of key digital library projects, and is being actively considered by others. A European interest group is currently in the early stages of creation, and an international editorial board has now been established to coordinate future developments.

Most projects currently considering using METS are still image and text archives, but active work is already underway, particularly at the Library of Congress, on using it for digital audio and video archives. The greatest risk to the higher education community deriving its full benefit from METS is the lack of standardization of metadata *content* which is needed to complement the standardization of *format* provided by METS: work needs to be undertaken to address this gap before its full benefits can be attained.

Keywords

Metadata;digital libraries;XML;standards.

Introduction: METS and metadata standards

Digital library technology has been well established for many years, and many major collections have converted their contents to electronic form. A vital component of a digital collection, as with its more traditional counterpart, is the provision of adequate information about its contents: this information, commonly referred to as metadata, allows users to find the material they require and assess its relevance, and allows the holding repository to administer it efficiently. As digital collections have proliferated, so have approaches to how their metadata should be handled.

Various attempts have been made over the last decade to bring some degree of standardization to digital library metadata. One key initiative in this area is the Dublin Core[1], a set of 15 very basic elements for describing electronic objects, which has been widely adopted as a generic interchange standard. Other projects have attempted to do the same for more specialized forms of electronic content, including the Open Archives Initiative[2], which is aimed at the e-print community, and the CEDARS project[3], which aims to provide a framework for information needed for the long term archiving and preservation of electronic collections.

Although each of these standards works perfectly well in the compass within which is it designed to function, what has been lacking until now has been any overarching framework within which *all* types of metadata for digital library objects can be contained. In the traditional library world, this need has been met for over 30 years by the MARC standard, which holds in a coherent structure all information needed to describe library materials: among other benefits, this allows the content of a MARC-encoded catalogue to be readily interchangeable with its counterparts elsewhere. The **Metadata Encoding and Transmission Standard (METS)**[4] attempts to fill this role in the area of the digital library, where the variety of metadata required is much greater: it acts as a *framework* within which the pre-existing metadata schemes described above, and any designed in the future, can be integrated in a logical and structured way, so allowing some of the advantages which the MARC standard gave the library world to be realized in its electronic equivalent.

Digital library metadata: the need for standards

METS is an newly emergent standard designed to encode all varieties of metadata necessary for a complete description of digital objects within an electronic library. Such objects may take the form of electronic texts, still images, digitized video, sound files or more interactive material such as VRML virtual environments. Until recently, no standardized method for encoding metadata on these objects has been available, and, as a consequence, every digital library project has tended to follow its own practice, often making use of whatever software package and data format its project team had become familiar with. As the number of such projects has proliferated, the need for a standard approach to metadata has become more acute: it is this need which the METS standard has attempted to address.

The Digital Library Federation[5], a consortium of libraries and related institutions which pioneers electronic library practices, has defined three key types of metadata necessary to describe a digital library object:-

Descriptive: information relating to the intellectual contents of the object, akin to much of the content of a standard catalogue record: this enables the user of a digital library to find the object and assess its relevance.

Administrative: information necessary for the manager of the electronic collection to administer the object, including information on intellectual property rights and technical information on the object and the files that comprise it.

Structural: information on how the individual components that make up the object relate to each other, including the order in which they should be presented to the user: for example, how the still image files that comprise a digitized version of a print volume should be ordered

The diversity of approaches taken by digital library projects until recently has had consequences which are far from ideal. In the world of standard library cataloguing, the universal application of the MARC standard has made the cross-searching of collections and the creation of large union catalogues commonplace – in the digital library, it remains very difficult to provide efficient searching across collections, and an onerous practice for projects to convert their records to new standards: cross-institutional collaboration on the creation of shared collections is, consequently, a difficult business. In addition, the practice followed by many projects to entrust their metadata to proprietary standards, often symbiotically tied to a given software package, stores up serious problems of conversion for those future times when the packages become obsolete.

Two types of standardization are required to allow allow the full benefits of interchangeability of metadata to be achieved: the standardization of the metadata containers themselves (the equivalent of the MARC format in the world of library cataloguing), and the standardization of the content to be held in these containers (the equivalent of cataloguing rules and practices in traditional library operations). The METS format is an attempt to provide the former, a standard, but flexible, format to hold the diverse metadata associated with a digital object in a form in which it can easily be shared, cross-searched, exchanged, searched and rendered for browsing and display purposes. The latter type of metadata, although outside the scope of this report, is equally important, and any implementation of METS must take into account the need for standardized cataloguing practices (including name authorities) if its full value is to be realized.

What METS does

METS is intended primarily as a flexible, but tightly structured, container for all metadata necessary to describe, navigate and maintain a digital object, primarily the three types defined by the Digital Library Federation. All metadata relating to a single object (which could be a single image, or all the components making up an item as complex as a digitized volume) are integrated into a single file: within this file, each type of metadata (descriptive, administrative and structural) is described in a separate section, which is linked to its counterparts by a comprehensive system of internal identifiers. The metadata itself may be held physically within the METS file, or may be held in external files and referenced from within the METS document: it may follow any preferred scheme, although a number of these are recommended specifically for use within METS.

METS is written in XML, a generic language designed for marking up electronic texts. XML is the most widely accepted interchange standard for the exchange of metadata currently available: its major advantages include the fact that it is not tied to any proprietary software platform, its robustness as an archival medium[6], and its ready interchangeability resulting from the fact that it uses standard ASCII code rather than a binary format to encode its data.

Until recently, all XML applications were defined in files known as DTDs: these *Document Type Definitions* contained the list of tags that were allowed within a given XML application, and rules dictating how they might be used (for instance, which tags were allowed within which). Within the last couple of years, a much more powerful standard for defining XML applications, known as *XML Schema*, has become available: this essentially uses a separate XML document to define the tags and rules for an XML application, and allows greater control of the content of these tags and how they are used. It is in this new standard that METS is written.

A METS file consists of five major sections, each describing a different facet of the digital object:-

Structural Map

The key part of a METS file, and the only compulsory section, is a description of the overall structure of the object (its structural metadata). This section describes the major components within the object, and how they relate to each other hierarchically. For example, if the object is a digitized book, this section will show

that the book as a whole is divided into separate chapters, and if these chapters themselves contain sections or subsections, it will show how these are nested together.

A very simple structural map for a book may look like this:-

```
<structMap TYPE="PHYSICAL">
     <div ID="title1.div.1"  LABEL="Chapter 1">
           <div ID="title.1.div.1.1" LABEL="Section 1">
                  <fptr FILEID="title1.image.1"/>
                  <fptr FILEID="title1.image.2"/>
           </div>
           <div ID="title.1.div.1.2" LABEL="Section 2">
                  <fptr FILEID="title1.image.3"/>
            </div>
       </div>
       <div ID="title1.div.2"  LABEL="Chapter 2">
           [Contents of Chapter 2 omitted]
       </div>
  </structMap>
```

This structural map shows that the volume is divided into two chapters, and that the first chapter is divided into two sections. Within these sections are pointers to the files that hold the images of the pages that make up this section: these are referenced by the FILEID attribute within the <fptr> (file pointer) element .

The structural map provides a logical layout for the structure of the whole object, and one that is easy to navigate using any XML-compatible software. It is straightforward, for instance, to write a stylesheet in the language XSLT (the language for rendering XML documents for display) with which the user can browse through the whole volume.

File Groups

The individual files of which the digital object as a whole is comprised are listed in the file group section. Each file is referenced by a file element, such as:-

```
<file ID="FILE003" MIMETYPE="image/jpeg">
          <Flocat LOCTYPE="URL">
               http://dlib.nyu.edu/tamwag/beame.jpg
          </FLocat>
</file>
```

The physical location of the file, usually given in the form of a URL, is indicated by the `<Flocat>` element, while its ID, by which it is referenced from the structural map, is declared by the ID attribute of the `<file>` element itself. Different manifestations of the same image, sound file or other digital file (for instance, thumbnails, archival copies, or delivery versions in the case of still images) may be grouped together in surrounding `<fileGrp>` elements, and so referenced together if necessary.

Descriptive Metadata

Descriptive metadata for a digital object, or any of its components, is held within a section of the METS file named <dmdSec>. Each <dmdSec> is labelled with a unique ID to allow it to be referenced from within the structural map. METS allows this metadata either to be held in external files which are referenced from within the METS file, or to be embedded directly with it.

In the former case, an <mdRef> element is used to point to the file containing the descriptive metadata, as in this example which points to metadata held in a file encoded in EAD:

```
<dmdSec ID="dmd001">
    <mdRef LOCTYPE="URN" MIMETYPE="application/xml" MDTYPE="EAD"
    LABEL="Berol Collection Finding Aid">urn:x-nyu:fales1735</mdRef>
</dmdSec>
```

The alternative approach allows the descriptive metadata to be embedded directly in the METS file, using an <mdWrap> element to contain it:

```
<dmdSec ID="dmd002">
    <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin  Core
Metadata">
        <dc:title>Alice's Adventures  in Wonderland</dc:title>
        <dc:creator>Lewis Carroll</dc:creator>
        <dc:date>between 1872 and  1890</dc:date>
        <dc:publisher>McCloughlin  Brothers</dc:publisher>
        <dc:type>text</dc:type>
    </mdWrap>
</dmdSec>
```

The above example shows some simple metadata for a digitized book encoded in Dublin Core. In fact, any metadata scheme may be used within an <mdWrap> element, but the METS editorial board suggest a number of these as particularly suitable, and recommend a number of XML schemas which can be used to extend METS to incorporate them. Currently four such schemas are available on the METS home page, including one for MARC records.

Administrative Metadata

Administrative metadata, including technical information on the digital files that comprise an object, and information on intellectual property rights attached to it, is dealt with in a manner analogous to descriptive metadata. <amdSec> elements, each identified by an ID, are used to record this metadata, which may be held in external files or embedded within the METS file using the <mdWrap> element. Any metadata scheme may be used for this, but the METS editorial board recommends particularly the NISO Technical Metadata for Digital Still Images Standard[7] for still images.

Behaviour

A late addition to the METS standard, the behaviour section allows information to be recorded on how components of the digital object should be rendered for the user: this may include information on specific software packages to be used, or on particular parameters to be used when rendering a file.

The importance of METS

Digital library technology is now well established, and the number of major collections coming online is increasing daily. Much time has been expended over the last decade in getting to grips with the requirements of imaging technology and discovering how it may sensibly be applied to library materials. Comparatively little attention, however, has been paid to the needs of metadata, an omission which has resulted in a confusing array of approaches to dealing with this vexed question.

Often digital library metadata is not up to the standard that is the norm in traditional library cataloguing: no agreed format, analogous to MARC, has been applied, and authority control for names and subject descriptions is often absent. The choice of technology used for digital library projects has often been dictated by the presence of an accepted software package within an institution, and the format of metadata within such projects has often been determined by the requirements of this software. In many cases, metadata is held in proprietary formats which in the case of older packages may no longer be readable.

METS has some immediate advantages over proprietary formats, especially binary ones, by virtue of being written in XML: it is archivally robust, human readable and readily interchangeable between XML-compatible applications. But this in itself is not sufficient: XML has been used by a number of digital library projects over the last few years, but the choice of DTDs used has been too diverse to suggest that agreement on a standard has been reached. Although many of these DTDs can be mapped to each other to some extent, which should render the interchange of their metadata easier, these mappings are almost always partial, and the steps that have to be negotiated to convert their metadata are far from simple. The need for a universally-applicable standard that will incorporate all key types of digital library object metadata is now paramount, and METS attempts to step into this role.

METS is well suited to fill this role for one key reason: unlike other XML applications which have been used as the basis of digital library projects (such as the EAD and TEI), it has been designed specifically for the various types of metadata that apply to digital library objects. It does not, therefore, require any adaptation or the use of tags for purposes removed from their originally intended function, as is often necessary when using pre-existing DTDs. It has a clearly-delineated internal structure which is easy to generate automatically from other applications (such as relational databases), although this this is at the expense of it being more verbose and less elegant than more traditional XML applications. Its ability to offer users a choice of embedding metadata directly within its structures or referencing externally held metadata gives it an extensive degree of flexibility and makes it relatively easy to use for the conversion of legacy data. It is also designed to be readily extensible, so that newly-enunciated requirements from the digital library community can be incorporated relatively quickly following a simple review process: it will, therefore, be able to grow to accommodate future changes in digital library practices.

Of course, the degree of flexibility provided by METS could also cause problems for its widespread applicability: a standardized container for metadata is of little value if the content of the metadata is itself not standardized to any degree. In the traditional library world, the MARC standard is complemented by the Anglo-American Cataloging Rules (AACR2) , which provides instructions for the content of catalogue records. METS in itself does not prescribe the content of the metadata it contains any more than the MARC standard does for its metadata.

The METS editorial board does itself attempt to address some of these problems by recommending a number of extension schemas for use with METS. These are simply external XML schemas which encode widely used metadata standards (such as Dublin Core), and can be incorporated into the METS schema by a commonly-used XML mechanism known as extension. Unfortunately, in many cases, these extensions are, like METS itself, no more than containers for metadata with convenient and standardized labels, and do not prescribe the contents of the metadata themselves.

To make effective use of METS and its associated schemas, digital library projects do need to adopt cataloguing rules governing the content of their metadata: AACR2 will in most cases suffice for this, as it

incorporates rules for electronic media. What is needed to allow METS to fulfill its potential as a standard akin to MARC in more traditional library practice is the widespread adoption of agreed cataloguing rules to complement the standardization it offers as a container for this content.

Products

As METS is an application of XML, which is designed specifically to be non-proprietary and so independent of any given software package, it can be implemented by any of a variety of XML-compliant products. To implement an XML-application usually requires an XML editor to create the documents themselves, a parser to check their validity (conformance to the DTD or schema), and an XSLT processor which allows them to be converted for display (usually in HTML). A wide variety of XML software is available to carry out these functions, much of it free and open-source[8].

However, owing to the fact that it is implemented in the relatively new format of XML-schema, the range of tools available that can handle this new application is relatively limited: the well-known XML-editor, XMetal[9], for instance, can read and edit XML-schema documents in its latest version (3.0), but not earlier ones. The widely-used XML parser, nsgmls[10], cannot yet handle XML schema, and so more modern parsers, such as Xerces[11], are needed to check the validity of METS documents. This situation is, however, improving all the time as XML software is revised to incorporate XML schema. For converting METS documents for presentation on the World-Wide Web, standard processors such as Saxon[12] can be used without modification.

METS is still a relatively new standard, but a number of XSLT stylesheets have already been made available to process METS documents for display. New York University, for instance, has devised a stylesheet to provide a page-turning application for bound volumes encoded in METS[13]. Other stylesheets have been published which allow metadata encoded in earlier DTDs to be converted to METS (such as the California Digital Library's stylesheet to convert from the MOA2 DTD to METS[14]). The number of these stylesheets is growing rapidly as more digital library projects adopt METS.

Other projects have been producing tools to assist in the administration of METS documents; of these, the most notable is Harvard University's METS Java Toolkit[15] produced by their Digital Library Initiative team. This is a package of software items which allows METS documents produced by various metadata providers to be integrated automatically into a central archive.

No commercially-produced turnkey digital library packages can accommodate METS at the time of writing, although several producers have an expressed interest in introducing METS capabilities into the next version of their products.

Current METS developments in HE and FE

The importance of the METS initiative has been recognized within the Higher and Further Education communities on both sides of the Atlantic, and several major projects are in the process of implementing it and developing it further. Oxford University's Digital Library[16], for instance, is using METS as the central component of its metadata system, and has established detailed procedures for mapping metadata from legacy digital projects to the scheme. The British Library and the National Library of Wales are both well advanced in plans to implement METS for their digital collections, and the former is trying to establish a Europe-wide interest group to promote its adoption.

The creation of a critical mass of METS-encoded metadata objects is not far away, and with this is coming its firm adoption as a standard. Some barriers, however, currently exist which impede its wider dissemination: these include poor documentation, a lack of support and a general lack of knowledge outside specialist circles of its existence and potential. Concertation activities will undoubtedly raise the standard's profile in the near future, however, especially if the proposed European interest group is successful. Further resources need to be allocated to improving support and documentation at the central METS website to make the task of implementation easier, especially for those not expert in XML. Fortunately an editorial

board has now been set up to oversee the future development of the standard and to rectify some of the deficiencies in the central resources currently available to support METS; these should facilitate greatly its wider adoption.

Assessment

METS is in its early stages but is already sufficiently well established amongst key digital library players that it can be reasonably be considered the only viable standard for digital library objects in the foreseeable future. As it is currently designed, METS can incorporate metadata for still images, text, moving images and digital audio, and it is in these areas that its initial applications are likely to be concentrated. Most METS projects currently underway are essentially collections of still images, although some also incorporate searchable electronic texts as complements to their image files: the Oxford Digital Library, for instance, will be using METS to provide integrated access to both types of digital resource.

Audio and video collections are fewer in number, but major collections have already announced their intention to use METS as the basis of their metadata in future developments. The Library of Congress' Digital Audio-Visual Preservation Prototyping Project[17], a five year project to design repositories for recorded sound and moving image collections, will be using METS as its primary metadata scheme. A definitive list of extension schemas for different classes of digital materials, which cover descriptive, technical and rights metadata, is being drawn up as part of this project: this should make the implementation of METS, which is its raw state is something of a skeletal framework, much easier, as it will provide well-designed components for each type of metadata which will slot easily into the overall structure.

By itself, METS offers no more than a coherent structure into which metadata can be put: it does not prescribe the content of the metadata itself. As already noted, rules on the content of metadata are just as important as the container that holds it, and the full potential of METS as a standard will only be realized when rules of this kind are adopted as widely and consistently as this container. Perhaps the greatest risk to the successful implementation of METS as is stands is that projects may invent adhoc localized solutions to the problems of metadata content, and so use the facilities of METS for the enhanced cross-searching of metadata that is not normalized sufficiently to allow this cross-searching to have much value.

In some ways the very flexibility of METS, which is its key strength when it comes to its incorporation into pre-existing projects with legacy metadata, is its greatest weakness: it is only part of the solution to the problem of standardizing digital library metadata. Its rapid adoption by the digital library community, and adequate support for this adoption, are both essential, but so are further developments, along the lines of the Digital Audio-Visual Preservation Prototyping Project outlined above, which will move the community towards a greater standardization of content. When both lines of development are fully implemented, a true standard for digital object metadata will have been achieved, and the academic community will at last be able to benefit from its adoption.

Glossary

AACR2: Anglo-American Cataloging Rules, the standard application guidelines for machine-readable cataloguing practice

ASCII: American Standard Code for Information Interchange – the most common format for text files on computers and the internet

attribute: within an XML document, a component of an element which modifies its meaning: for example: type="PHYSICAL"

DTD: Document Type Definition

Dublin Core: a basic set of 15 metadata element designed to represent core fields for the description of any electronic resource

EAD: Encoded Archival Description, an XML standard for the encoding of archival finding aids

element: the tags within an XML document, corresponding to the fields of a database: they have the format <tag>..</tag>**,** for instance <structMap>....</structMap>

e-print: electronic versions of scholarly papers intended to increase access to scholarly research

MARC: MAchine Readable Cataloguing: the standard within traditional libraries for cataloguing information

METS: Metadata Encoding and Transmission Standard

MOA2 : Making of America II: a DTD which attempted to encode digital library metadata, now superseded by METS

NISO: National Information Standards Organization

open-source: software released (usually free-of-charge) with its source code under a licence which allows it to be modified by users

TEI: Text Encoding Initiative: a widely used XML application for electronic text

URL: Universal Resource Locator: the "address" of an internet resources (for example, http://www.loc.gov/)

VRML: Virtual Reality Markup Language

XML: eXtensible Markup Language

XML parser: a software package used to validate an XML-encoded document (ie. check that it conforms to the DTD or schema in which it is encoded)

XSLT: eXtensible Stylesheet Language Transformations – a language used to write stylesheets for rendering XML documents for display

References

1. http://www.dublincore.org
2. http://www.openarchives.org
3. http://www.leeds.ac.uk/cedars
4. http://www.loc.gov/standards/mets
5. http://www.diglib.org/dlfhomepage.htm
6. Coleman, James and Willis, Don, *SGML As a Framework for Digital Preservation and Access,* Council on Library and Information Resources, 1997 (ISBN: 1887334548) for a description of the archival value of SGML/XML
7. http://www.loc.gov/standards/mix/
8. See http://www.oasis-open.org/cover/publicSW.html for an extensive list of publicly available XML software.
9. http://www.sq.com
10. http://www.jclark.com/sp/nsgmls.htm
11. http://xml.apache.org/xerces2-j/index.html
12. http://saxon.sourceforge.net/
13. http://dlib.nyu.edu/metstest/
14. http://sunsite.berkeley.edu/mets/moa2mets/
15. http://hul.harvard.edu/mets/
16. http://www.odl.ox.ac.uk/

§http://lcweb.loc.gov/rr/mopic/avprot/avprhome.html