

TSW 03-06
October 2003

MODS: Metadata Object Description Schema

Richard Gartner

Pearson New Media Librarian
Oxford University Library Services

Copyright JISC 2003

Executive Summary

As digital library projects have expanded in number, the need for standardization of their metadata has become more acute. A standard *framework* for this metadata is now provided by METS (Metadata Encoding and Transmission Standard), but its usefulness is limited by the lack of a satisfactory standard for metadata *content*. In the area of descriptive metadata, the key initiative until now has been Dublin Core: it has been hampered, however, by the conflicting needs of interoperability and precision which it has failed to resolve. MODS (Metadata Object Description Schema), a new standard published by the Library of Congress' Network Development and MARC Standards Office, aims to allow reconcile these demands in a single, coherent scheme for describing digital objects.

MODS, written in XML, provides 19-top level elements for describing objects, and a further 64 sub-elements under these. These cover standard bibliographic facets such as titles, names of creators and contributors, subject and classification numbers, and also such areas as physical description, information on access restrictions, and genre. It also includes mechanisms for extending its element set by incorporating other XML documents within its structure, and allows records for related objects to be incorporated in this way. In addition, it includes facilities for authority control, and is fully granular in the level of descriptions it provides.

MODS has arrived at a time when the current heterogeneity of approaches to metadata is becoming untenable: it reconciles the problems Dublin Core has experienced by providing a richer element set but still allowing extensibility when required. As it is an XML application, it is non-proprietary and not tied to any given software package. Tools have already been written to convert MARCXML records to MODS and further similar tools are easily written.

MODS has already been used for a wide variety of classes of material, from books to multi-media, and has been adopted by a number of major digital library projects. Its use is also being widely promoted by key bodies in the digital library world. When used in conjunction with other key standards, such as METS, it offers the potential for a fully coherent, integrated metadata strategy which will much enhance access to digital materials worldwide. However, until a critical mass of MODS records is created, the full advantages of its adoption, in terms of facilitating the creation of collections which can easily interact with each other, will not be obvious: its potential impact is therefore long-term. To fully realize this potential will require its widespread adoption coupled with a greater degree of rigour in standardizing metadata content than has often been the case before now.

Keywords

Metadata; descriptive metadata; digital libraries; XML; standards.

Introduction: standards for descriptive metadata

Since the library community first became interested in the digitization of its collections, the technologies involved in the conversion of materials to digital form have become well established and the number of digital library projects has expanded exponentially. As the volume of material being made available digitally has increased, however, so has the need to standardize the metadata used to describe and administer it. No standards have yet become fully entrenched in the area of metadata, and the approaches being taken towards it are diverse: the downside of this variety, unfortunately, is that few of the benefits from standardization to which the traditional library world has grown accustomed, particularly the sharing of records and the cross-searching facilities which allow union catalogues to be created, have accrued in the digital arena. This state of affairs is becoming increasingly untenable as the mass of digital materials increases.

The problem of providing a standardized *framework* for digital library metadata has been tackled by the creation of the METS (Metadata Encoding and Transmission Standard) standard, which is covered in an earlier TSW report¹. This fulfills a similar function to the MARC record in a traditional library catalogue by providing a standardized container within which digital object metadata may be held. It does not, however, prescribe the form of *content* which should be used within its structure to describe digital library objects. Without such a standardization of content, the potential of METS to allow the interchange and sharing of metadata can only be half-realized.

In the area of descriptive metadata, the intellectual content of an item which is necessary to find it and judge its usefulness for a given user's requirements, the key attempt at standardization in recent years has been the *Dublin Core* (DC)² initiative. This defined 15 key elements³ for describing digital resources, and a set of mechanisms to qualify these for those cases when more detailed definitions are required. Since its inception in 1995, DC has proved enormously influential: it has been adopted by a vast array of online projects⁴ and has become the closest that the digital library community has yet come to a fully accepted metadata standard.

Unfortunately, the dual approach to semantic breadth (the use of simple fields and qualifiers to refine them) that DC has taken has reduced its value as either a metadata container or as a medium of exchange. The 15 elements by themselves are often found to be too broad to allow their unqualified use: for example, the DC field *creator* can cover a large array of individuals or organisations responsible for the creation of an object, from authors to editors to compilers and many others. Concatenating all of these into one field is often inadequate for a useful bibliographic description. On the other hand, qualifying elements to distinguish, for example, authors from editors in the *creator* field reduces the interoperability of the DC record, so negating one of the key *raison d'être* of the standard.

Other important metadata schemes have been published for narrower classes of material: these include ONIX⁵, an XML scheme produced by the book trade to allow publishers to exchange bibliographic information, CDWA⁶, a set of elements for describing works of visual art, and IMS⁷, an XML application for the description and management of online teaching materials. All are well suited for their given applications, but of little use outside: despite the possibility of producing mappings of their constituent fields to each other, the degree of exchangeability of metadata recorded in these differing standards remains poor, and the potential of standardized metadata to make the pooling or exchange of digital materials easier remains generally unrealized.

This report introduces *MODS* (*Metadata Object Description Schema*)⁸, a new metadata standard published by the Library of Congress' Network Development and MARC Standards Office (the unit responsible for the maintenance of several key bibliographic standards, including MARC and METS). It aims essentially to reconcile the dual demands of interoperability and precision which have reduced the usefulness of Dublin Core and the other initiatives mentioned above: it tries to do this by providing a more extensive element set than unqualified Dublin Core (so allowing the vast majority of fields necessary for most bibliographic descriptions to be incorporated into its standard framework), while still allowing extensibility for the more specialized components of a description which are of less value for interoperability. By doing so, it should make access to digital materials, and the pooling of digital resources in general, much easier to achieve.

How MODS works

MODS attempts to reconcile the conflicting demands of breadth and specificity which have affected in particular Dublin Core and hindered its applicability as a generic medium for metadata: although, like that standard, it does contain a mechanism for extension, it offers a much more detailed element set than unqualified DC, so allowing a much large number of items to be described without resorting to using elements outside its core set. It is therefore much more readily interchangeable with other MODS records, and also, via detailed mappings, to other metadata schemes.

MODS, like METS and several other metadata standards, is written in XML (eXtensible Markup Language), which ensures that it is independent of any software package and is archivally robust. It is based on a subset of MARC, but unlike this standard uses English-language rather than numeric field names, so that a knowledge of the MARC standard and its conventions is not required by the MODS user. MODS is also easier to use than MARC as it has cut down the standard to a smaller element set and has reorganized its constituent fields, often re-grouping them into more logical components.

When compared to Dublin Core, MODS offers a much fuller and more specific element set: 19 top-level elements and a further 64 under these allow detailed descriptive records to be created for many classes of material. The MODS website provides on its home page sample files for 9 such classes, including books, serials, maps, sound, video, and mixed media, amply illustrating the wide range of objects that can be handled by the standard.

This is a sample MODS file for a book, taken from the Library of Congress' MODS website:-

```
<mods>
<titleInfo>
  <title>Sound and fury :</title>
  <subTitle>the making of the punditocracy </subTitle>
</titleInfo>
<name type="personal">
  <namePart>Alterman, Eric</namePart>
  <role>
    <text>creator</text>
  </role>
</name>
<typeOfResource>text</typeOfResource>
<genre authority="marc">bibliography</genre>
<originInfo>
  <place>
    <code authority="marc">nyu</code>
    <text>Ithaca, N.Y.</text>
  </place>
  <publisher>Cornell University Press</publisher>
  <dateIssued>c1999</dateIssued>
  <dateIssued encoding="marc">1999</dateIssued>
  <issuance>monographic</issuance>
</originInfo>
<language authority="iso639-2b">eng</language>
<physicalDescription>
  <form authority="marcform">print</form>
  <extent>vii, 322 p. ; 23 cm.</extent>
</physicalDescription>
<note type="statement of responsibility">Eric Alterman.</note>
<note>
  Includes bibliographical references (p. 291-312) and index.
</note>
```

```

<subject authority="lcsh">
  <topic>Journalism</topic>
  <topic>Political aspects</topic>
  <geographic>United States.</geographic>
</subject>

<subject authority="lcsh">
  <geographic>United States</geographic>
  <topic>Politics and government</topic>
  <temporal>20th century.</temporal>
</subject>

<subject authority="lcsh">
  <topic>Mass media</topic>
  <topic>Political aspects</topic>
  <geographic>United States.</geographic>
</subject>

<subject authority="lcsh">
  <topic>Television and politics</topic>
  <geographic>United States.</geographic>
</subject>

<subject authority="lcsh">
  <topic>Press and politics</topic>
  <geographic>United States.</geographic>
</subject>

<subject authority="lcsh">
  <topic>Talk shows</topic>
  <geographic>United States.</geographic>
</subject>

<classification authority="lcc">PN4888.P6 A48 1999</classification>
<classification edition="21" authority="ddc">071/.3</classification>
<identifier type="isbn">0801486394 (pbk. : acid-free, recycled paper)</identifier>
<identifier type="lccn">99042030</identifier>

<recordInfo>
  <recordContentSource>DLC</recordContentSource>
  <recordCreationDate encoding="marc">990730</recordCreationDate>
  <recordChangeDate encoding="iso8601">20000406144503.0</recordChangeDate>
  <recordIdentifier>11761548</recordIdentifier>
</recordInfo>

</mods>

```

MODS provides 19 top level elements (all but one of which are optional) to describe the broadest facets of an object. As is shown in the above example, the scope and content of each of these top-level elements is generally obvious from their names. Many are wrapper elements which are subdivided into narrower subelements with a more specific semantic scope: others are unstructured internally and so contain essentially free-text. A full explanation of all MODS elements and their usage is available at <http://www.loc.gov/standards/mods/mods-userguide-elements.html> - the following attempts to give a brief overview of the most important of these.

titleInfo

The only compulsory top-level element, **titleInfo**, is a wrapper for 5 possible sub-elements, of which one, **title**, is itself mandatory: **title** is obviously used to record the main title for the item, and its sibling elements can record such components as a sub-title, or, if the object is part of a larger work (an *analytic*), to note the title and other details of its parent (*monograph*) item. A final element, **nonSort** specifies those parts of a title which should be counted as non-filing characters when building indexes.

titleInfo may be qualified by a number of attributes which, amongst other things, can indicate what type of title is being recorded (for instance, a uniform or parallel title), the language in which it is recorded, how it has been transliterated if it was originally given in another script, and what form of authority control has been used for it.

name

The equivalent of the *Creator* and *Contributor* fields in Dublin Core, **name** is used to record people or organisations responsible for the creation of the intellectual content of the item, or to record those who contributed in some way to its creation (for instance, illustrators or printers). A **type** attribute indicates whether the name is a personal or corporate one, or, if the item is a set of conference proceedings, records the name of a conference.

Names may be split up into their components (family names, given names) and also recorded in an unstructured form for display purposes. The role of the person or organisation in the creation of the object may be indicated in several ways, including relator codes (such as the MARC codes⁹), or in plain text. A textual description may also be provided to describe the person or organisation in more detail, a feature not available in MARC.

originInfo

originInfo is another wrapper element, which brings together information on the provenance or publication of the item. Subelements record the date of origin of the item, which can be its date of publication, creation (in the case of unpublished materials or manuscripts), or capture in the case of surrogates of original items. This section also records details of the publisher of the item, and whether the item is monographic or continuing, and, in the latter case, its frequency of issue. The range of metadata that can be recorded in this section is extensive and very flexible, so allowing it to describe items in almost any medium.

physicalDescription

Another wrapper element, **physicalDescription**, contains a variety of sub-elements which allow a basic description of the object's physical characteristics. Many of these are relevant only to electronic resources: they include **internetMediaType**, which records the format of the data described (usually given in the form of a MIME type such as "text/html"), **reformattingQuality**, an indication of the quality (in terms of resolution and bit-depth) at which the item was scanned, and **digitalOrigin**, which records whether the object was born-digital or was reformatted from an original item in another medium.

More traditional media can be described in relatively limited ways only, principally by means of an **extent** element, which records the number of pages, illustrations etc, and by a **note** element, which can hold unstructured information on the physical characteristics of the object.

subject

subject is a further wrapper element used to describe the intellectual content of the item by means of subject terms taken from any desired taxonomy. The **subject** element sub-divides into components covering differing types of subject terms, such as names, geographic terms, or temporal ranges. One sub-element, **hierarchicalGeographic** can define a hierarchy of geographic terms, allowing browsing from the more general (such as continent) to the more specific (such as city). Another sub-element, **cartographics**, allows the detailed recording of spatial coordinates, in addition to the scale and projection used in maps. In sum, MODS provides a rich and highly flexible set of terms for subject descriptions and allows the incorporation of any number of taxonomies.

relatedItem

A very useful element in the context of a collection of items which share some interrelationship (such as a collection of digitized articles from a journal), **relatedItem** allows full MODS records for related items to be

embedded within its enclosing tags. A **type** attribute with a closed list of values (including *preceding*, *succeeding*, *original* and *constituent*) specifies the type of relationship. This element carries out the same function as the **Relation** in Dublin Core, but is much more flexible in its usage.

extension

Although MODS offers a much more extensive element set than unqualified Dublin Core, it is still possible that it may not satisfy all the metadata requirements for a given object. In such cases, it offers the facility to extend its element set by allowing metadata recorded in alternative schemes to be embedded within a MODS record. This additional metadata is delineated by a different XML namespace (a mechanism for distinguishing elements which may otherwise have the same name and so allowing them to be included in the same document). This may, for instance, be used in records for manuscripts to allow elements from a scheme such as MASTER¹⁰ to supplement those in MODS itself.

Other MODS elements

The remaining top-level MODS elements are self-explanatory:-

typeOfResource: the type of object being recorded, eg. text, cartographic, multimedia. The terms used are taken from a closed list

genre: a more specific term than typeOfResource, this allows detailed genre terms (taken from any approved source) to characterize the item

language: records the language of the item as a whole, using a code provided by one of two approved authority lists

abstract: a description of the intellectual content of the item or a link to a description of this kind

tableOfContents: a listing of the contents of the item, which may either be recorded explicitly or provided as a link to such a list

targetAudience: a term for the intended audience (eg. adult, adolescent) for the item, preferably taken from the controlled MARC list (available at <http://www.loc.gov/marc/sourcecode/target/targetlist.html>)

note: a "catch-all" element for recording information not suited to any other element

classification: the classification number for a resource under an approved scheme such as Library of Congress Subject Headings or the Dewey Decimal Classification

identifier: a unique number or code assigned in accordance with an approved scheme, such as the ISBN or ISSN number for a monograph or periodical respectively

location: a record of the physical location of the item, including the repository and the item's shelfmark

accessRestriction: information on how access to the item is restricted, including copyright information

recordInfo: a wrapper element for information on the creation of the MODS record itself, including the date of creation, control numbers etc.

Other features of MODS

All MODS elements whose content can be subject to a controlled sets of terms can use an **authority** attribute to give the name of the scheme from which these terms are taken: for example, the **authority** attribute for the **role** element is set to "marcrelator" to indicate that the code indicating the role of the person or institution is taken from the MARC Relator codes list. The use of controlled vocabulary from an authoritative list is recommended whenever possible to make interchangeability and cross-searching easier.

One other key feature of MODS is its granularity: it can be used to describe either an item as a whole, or any sub-components of an item. An electronic version of a journal, for example, may use MODS fields to record information on the journal itself, and then nested MODS records to describe each constituent article. This is a major advantage over Dublin Core which can only replicate this feature much more crudely with its **relation** field.

MODS has an extensive system of IDs on many elements to allow detailed cross-linking within a record, and also to allow a particular element to be referenced from outside the MODS record. It also has the attribute **xlink:href** available on many elements to allow them to reference external files holding their content - for example, an abstract may be held in an external file and referenced from within the MODS file in this way:-

```
<abstract xlink:href="www.myserver.com/abstracts/abstract1.html" />
```

Detailed and intricate linking between MODS files and other XML applications is therefore readily supported within the MODS schema.

Why MODS matters

MODS is a timely development as it has arrived when it has become increasingly apparent that the heterogeneity of approaches to metadata for objects within a digital library has become untenable for the reasons given in the introduction to this report. It fills the gap left previously by the absence of a useable standard for descriptive metadata and offers the potential, with other standards such as METS, of offering a fully standardized environment for holding all digital library metadata.

MODS arrives when the failings of Dublin Core as a solution to the problem of standardizing descriptive metadata have become apparent, and offers an approach which reconciles reasonably successfully the divergent demands of interchangeability and precision which have caused problems for DC. In particular, it offers a richer set of elements than unqualified Dublin Core, which is the only form in which DC is truly interchangeable: it therefore allows much more precise cross-linking and searching between records without losing ready exchangeability. It is also much more controlled than qualified DC, allowing almost all of the functionality that projects have sought in the past by introducing qualifications to DC, but doing so in an organized and authoritative manner which does not compromise the interchangeability of their records.

MODS will realize its full potential when used in a unified XML environment for metadata for digital objects: in particular, it forms a useful adjunct to METS, and allows it to realize its full potential as a medium for enhancing the interchangeability of records by addressing the question of standardizing the *content* of digital library metadata. When used in conjunction, the two make the exchange of digital library metadata much easier and bring some of the advantages that standardization has introduced to the traditional library world.

Products

MODS is an XML application, and so specifically designed to be non-proprietary and not tied to any given software package. Which tools are used to create and process MODS records will depend of the specific

environment in which it is to be implemented. MODS records may be created manually with any XML editor or even a standard text editor: they may also be generated from database packages or created from pre-existing MARC records. Any process that is currently in use to export database records to a text-based format may be used to produce MODS records, although, obviously, this process will be easier in a system which already includes an XML export function.

One tool designed specifically for the creation of MODS documents is a stylesheet to produce MODS records from those encoded in the MARCXML scheme¹¹: this will allow the easy bulk conversion of current records which are either encoded in this scheme, or can be generated in this format from a pre-existing online catalogue.

Current MODS developments in HE and FE

Although MODS has only been available since June 2002, it has already been adopted by a number of important projects and its wider use is being strongly advocated by a number of key authorities. Some key projects which have already adopted MODS include BiblioVault¹², a digital repository for scholarly books maintained by the University of Chicago Press, Chopin Early Editions¹³, a collection of early printed editions of Chopin scores also at the University of Chicago, and the Electronic Boethius Project¹⁴ at the University of Kentucky. Other important projects which currently use Dublin Core (such as Oxford University's Digital Library¹⁵) are actively considering moving to MODS.

As with any new standard, the main barrier to the success of MODS is the current lack of a critical mass of encoded material large enough for its advantages as an interchange medium to become apparent. Projects adopting MODS are making a valid investment in the future but, without such a mass, will not have any obvious returns in the short run to make its advantages apparent. However, MODS does benefit from its provenance as a creation of the Network Development and MARC Standards Office at the Library of Congress, which will ensure its wide dissemination in the appropriate communities. It is also well supported by documentation at its main website and has a flatter learning curve associated with it than other new standards such as METS. Wider concertation activities, including training workshops, will be necessary to ensure its wider implementation and the realization of its potential. This will require further efforts by its creators and its early implementors to publicize the standard and the advantages accruing from its adoption.

Assessment

MODS succeeds to a large extent in reconciling the conflicting demands of interoperability and precision which have caused problems for other metadata schemes: it offers a rich but prescriptive set of core elements, and an extensibility mechanism to meet any more specific requirements for a full description of a given item. It provides full granularity, making it suitable for the description of each component of a complex item at the level it demands, and is designed to integrate well with other pre-existing standards (most obviously MARC21).

MODS goes a substantial way towards providing a standardized environment for library metadata, but it should be only one component in the overall metadata strategy for a digital library: to realize its full potential, it needs to be used in conjunction with other schemes covering other facets of a complete metadata system. For instance, it meshes very well with METS, which can provide an overall framework within which all the metadata for a digital object can be contained and whose editorial board recommends it as an approved schema. It can also be used in conjunction with any standard for administrative metadata, the information needed to manage a digital object. Used with these robust standards, the full potential of MODS for enhancing access to digital materials should become more fully realized.

In many ways MODS is a "behind-the-scenes" technology, and its potential impact is long-term: the adoption of MODS will not be immediately apparent to the user of a digital library system, but over a longer timespan, its ability to make it easier for digital collections to interact with each other will produce tangible benefits. The lead-in time for new projects should shorten considerably as they will not have to start from scratch devising

their own schemes for metadata content: nor need they spend time and effort adapting Dublin Core or similar schemes to their requirements. Disparate collections located physically apart should be able to link up more readily and with less effort spent on technical development than is necessary at present. Union collections, similar to the large union catalogues that are such a key feature of the traditional library environment, should become possible with less investment in time and resources than is currently required to map disparate metadata contents to each other.

Realizing the potential of MODS will, however, require more than its widescale adoption: the content of the metadata it holds must also be standardized and rationalized. The traditional library catalogue has long adopted cataloguing rules to ensure the consistency of its records and the same degree of discipline needs to be adopted when describing digital objects if the benefits of standardization are to be applied. To allow interaction with traditional materials, the "hybrid library" to which so much attention is currently being paid within the library community, the same cataloguing rules used there, or at least ones which are fully compatible with them, need to apply within the digital library. Technology can compensate for some irregularities in metadata content, for example by allowing fuzzy searching, but this is a poor substitute for consistent original records and can never match such a record for precision. To use MODS to its full potential will, therefore, require attention to these basic issues which have long since been tackled in the traditional library.

Glossary

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

administrative metadata: information necessary for the management of an item in a repository, including its technical specification and information on access rights

analytic: a bibliographic item which forms part of a larger unit (for instance, a journal article)

authority control: the process of verifying entries in an online catalogue and ensuring that all entries for the same name and title are located together

CDWA: Categories for the Description of Works of Art

descriptive metadata: information on the intellectual content of an item, analogous to the main part of a traditional library catalogue record

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

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

element: the tagged components with an XML document, which correspond to the fields of a database: for instance, <titleInfo>

IMS: Instructional Management Systems Metadata

ISBN: International Standard Book Number

ISSN: International Standard Serial Number

MARC: MACHine Readable Cataloguing: the standard used within libraries for cataloguing information

MASTER: Manuscript Access through Standards for Electronic Records: an XML application for recording detailed descriptions of medieval manuscripts

METS: Metadata Encoding and Transmission Standard

MODS: Metadata Object Description Schema

MIME type: an identifier used to uniquely define different file types: eg "text/html"

analytic: a unit which may be subdivided into smaller units (known as analytic units)

namespace: a mechanism for allowing XML elements with the same name to be included in the same document by prefixing them with a code - for example, mets:title and mods:title

parallel title: the title of a work in another language from that of its original main title

ONIX: Online Information eXchange

relator codes: codes to define the nature of the relationship of a named person or organisation to a library item:
eg prt=printer

uniform title: a title by which a work is to be identified for cataloguing purposes, to ensure that versions of the same work with different titles on their title pages are located together

XML: eXtensible Markup Language

References

1. http://www.jisc.ac.uk/index.cfm?name=techwatch_report_0205
2. <http://www.dublincore.org>
3. <http://www.dublincore.org/documents/dces/>
4. A list of many of these may be found on the DC website at <http://www.dublincore.org/projects/>
5. <http://www.editeur.org>
6. <http://www.getty.edu/research/institute/standards/cdwa/index.html>
7. <http://www.imsproject.org/metadata/>
8. <http://www.loc.gov/standards/mods/>
9. <http://www.loc.gov/marc/relators/>
10. <http://www.cta.dmu.ac.uk/projects/master/>
11. <http://www.loc.gov/standards/marcxml/xslt/MARC21slim2MODS.xsl>
12. <http://www.bibliovault.org>
13. <http://chopin.lib.uchicago.edu/>
14. <http://beowulf.engl.uky.edu/~kiernan/eBoethius/mainpage.html>
15. <http://www.odl.ox.ac.uk/>