9th International Symposium on Electronic Theses and Dissertations
IXe Symposium international sur les thèses et mémoires électroniques
June 7 – 10 Juin, 2006, Quebec City / Québec, Canada

MODS Meets Manakin: Innovations in the Texas Digital Library's Thesis and Dissertation Collection

Brian E. Surratt
Fondren Library, Rice University, Houston, Texas, USA
besurratt@rice.edu

ABSTRACT

The Metadata Working Group (MWG) of the Texas Digital Library (TDL) was charged with developing a common descriptive metadata standard for electronic theses and dissertations (ETDs) by December 2005. Based on the needs of the TDL and the characteristics of the existing collections, the MDW adopted the Metadata Object Description Schema (MODS) and subsequently developed an application profile tailored to ETDs. This paper describes the decisions the MWG made in order to express the elements of the ETD Metadata Schema (ETD-MS) in MODS. The paper also discusses the TDL's use of Manakin, a module for DSpace that allows the development of a custom user interface based on XML, and how Manakin interacts with MODS metadata.

Keywords: Electronic theses and dissertations, ETD, Metadata Object Description Schema, MODS, Manakin

1. INTRODUCTION

The Texas Digital Library (TDL), formed in 2005, is a consortium of the five members of the Association of Research Libraries in Texas. The members are the University of Texas, Texas A&M University, the University of Houston, Texas Tech University, and Rice University. The mission of the TDL is to "provide a digital infrastructure for the scholarly activities of Texas universities." It is unique in that it includes four large public university systems as well as one private research university.

The first project of the TDL was to develop a common repository of the electronic theses and dissertations (ETDs) published by the member libraries. At the time the project was initiated, two of the member universities, the University of Texas and Texas A&M, had ETD collections. The University of Texas provides access to the local ETD collection through MARC records in the library's catalog and Texas A&M provides access through Dublin Core records in DSpace. The Metadata Working Group (MWG) of TDL was charged with developing a common metadata standard that would allow members to share metadata in the TDL repository.

2. CHARACTERISTICS OF MARC AND DUBLIN CORE

In the United States, universities have traditionally provided access to theses and dissertations through the library catalog. Since the advent of online public access catalogs (or OPACS), this has meant applying the Anglo American Cataloging Rules (AACR2) and encoding the bibliographic information in a Machine Readable Catalog (MARC) record. A recent report on the cataloging treatment of theses and dissertations in the United States (Hoover and Wolverton, 2003) only acknowledges MARC cataloging and is essentially blind to other formats and standards, including Dublin Core and ETD-MS.

MARC cataloging has a rich history. Traditional OPACS have strengths, such as authority control for names, but it is becoming clear that AACR2 and MARC cataloging are not suited to meet the

descriptive metadata needs of ETDs. Developed in the 1960s, MARC is based on print catalog cards. It is a legacy standard and its shortcomings are magnified in the web environment. As a syntax for encoding bibliographic information, its characteristics are idiosyncratic (eg. fixed fields, indicators, etc.) especially compared to modern markup languages that were purpose-built for the web. Because it is so closely associated with AACR2 and ISBD punctuation, it mixes metadata content with formatting. Our OPACs, which are our primary systems for managing MARC records, are not designed to store content, but rather to point to the content in some other location. Our digital library systems, on the other hand, are not designed to store MARC records with our digital content. When libraries do use MARC to describe ETDs, it is applied inconsistently. Hoover and Wolverton (2003) demonstrated that MARC cataloging is inconsistent for information such as the genre, discipline, thesis advisor, and subjects. The ETD community has expressed the need for much of this information (see for example the ETD-MS, and Frodl and Korb (2005)), but the major cataloging guidelines either do not address how to encode this information or provide vague and inconsistent advice. AACR2 does not address the issue of advisor names. *Bibliographic Formats and Standards* states that "added entries for advisors, the institution, made-up thesis collections or series titles" should be placed in locally defined fields, which encourages local policies and discourages establishing common standards.

ETD-MS, the descriptive metadata standard developed by the NDLTD, is valuable for defining the information that we want to know about theses, but because it is expressed through Dublin Core, it inherits Dublin Core's flaws. Guenther (2003) points out many of these flaws. There are not a sufficient number of elements to describe digital resources and the existing elements are too broadly defined. Dublin Core lacks a sufficient structure. It does not specify a syntax, so its implementation in networked systems is inconsistent. It does not possess what Guenther calls a substructure, or a hierarchy of elements, so there are problems with describing component parts, relating elements to component parts, relating descriptive elements to each other, and providing attributes and qualifiers to individual elements. These problems emerge distinctly when Dublin Core is applied to ETDs.

The ETD-MS field "dc.contributor.role" serves as an excellent example. As written, this field is impractical to implement. According to Dublin Core, elements should be represented as strict name-value pairs:

dc.contributor=Smith, John
dc.contributor.role=author
dc.contributor=Brown, Jane
dc.contributor=advisor

No order for elements is specified in Dublin Core, hence…

dc.contributor=Smith, John
dc.contributor=Brown, Jane
dc.contributor=advisor
dc.contributor.role=author

…is equally valid. If this is done, Dublin Core does not provide a way to relate the two fields together in the description. If, alternatively, the role is expressed as an attribute of the contributor field, the way it is done in the ETD-MS example on the NDLTD site…

<contributor role="chair">Joseph W. Roggenbuck</contributor>

…then it is not valid Dublin Core and hence dose not follow any particular standard. Similar problems emerge when Dublin Core is used to represent various dates in the life of the ETD, as well as descriptions of component parts of ETDs. Anecdotal evidence has shown that the application of Dublin Core to ETDs has been inconsistent. Dublin Core is designed for cross-domain (ie. general) use, but institutions that host ETD collections often have specialized needs. MODS is better suited to address those needs.

3. TDL'S MODS APPLICATION PROFILE

The TDL found that MODS provides advantages over both MARC and Dublin Core. Developed by the Library of Congress in 2002 (Guenther 2003), it is partly based on MARC and carries forward the best features of this traditional cataloging format in a syntax that is defined in XML. Furthermore, it avoids many of the problems associated with Dublin Core. We were able to express almost all of the elements of ETD-MS in a satisfactory manner. For those elements that we could not express in MODS, MODS provides a <mods:extension> element that allowed us to refer directly to the ETD-MS schema. Our philosophy was to define and specify the use of mandatory elements and significant optional elements. We also allow the use of any valid MODS element, even if it is not included in our application profile.

Title information is encoded in the <mods:titleInfo> wrapper element. The <mods:title>element is mandatory and the <mods:subtitle> element is optional. The name of the author is encoded in a <mods:name> wrapper element with the type attribute set to "personal." The given name and family name are encoded in <mods:namePart> subelements. The birth date is optional. If it is included, it is placed in a <mods:namePart> subelement with the type attribute set to date. The name of the advisor is a mandatory element and is encoded in a <mods:name> wrapper element. The role of each name is specified using the <mods:role> wrapper element. The MARC relator terms "Author" and "Thesis advisor" are used to define roles. The names of committee members are optional.

Two dates are encoded in the <mods:originInfo> wrapper element: the creation date and publication date. The creation date is defined as "the date the student graduates or the date the degree is conferred" and is encoded the <mods:dateCreated> subelement. The publication date is defined as "the date the ETD is released to the public" and is encoded in the <mods:dateIssued> subelement.

The type of resource is mandatory. The <mods:typeOfResource> element is generally equivalent to the MARC leader/06. The values for this field come from a controlled list defined in MARC that includes "text," "sound recording," "moving image," and "software, multimedia" among others. Genre is mandatory and is encoded in the <mods:genre> element. All ETDs, regardless of level, are encoded with the MARC genre term "theses." Information regarding the physical details of the ETD is mandatory and is encoded in the <mods:physicalDescription> element. The MARC format term "electronic" is encoded in the <mods:form> element. The MIME type is encoded in the <mods:internetMediaType> element. MODS also has field to record whether the item described is born digital or reformatted into a digital format and this is recorded in a <mods:digitalOrigin> element.

"Langauge" is a mandatory element in the record, and can also be used as an attribute of any other MODS element. Other mandatory elements include abstracts, subjects, and information about the MODS record itself.

ETD-MS also includes extension elements outside of Dublin Core: thesis.degree.level, thesis.degree.discipline, and thesis.degree.grantor. TDL made all of these elements mandatory. The degree granting institution is represented in a <mods:name> element with the type attribute set to "corporate" and "Degree grantor" in the <mods:roleTerm> element. Degre level and degree discipline are not defined in MODS, so in order to include these, we had to use the <mods:extension> element and refer back to the ETD-MS.

4. OUTSTANDING ISSUES FOR MODS METADATA

There are a number of outstanding issues that the MWG will continue to investigate. The application profile does not include two elements specified by ETD-MS, publisher and rights. Traditionally, print

theses and dissertations have been unpublished manuscripts, so library catalog records do not include publisher information in MARC records. For this reason, one of our member libraries did not want to include this element. In fact, cataloging rules are inconsistent on this topic. Section 3.1 of *Bibliographic Formats and Standards* states that theses and dissertations are usually considered unpublished manuscripts, but AACR2 rule 9.4B2 says that remote access electronic resources should be considered published. Libraries seem to have a hard time thinking of themselves as publishers. Furthermore, the issue of rights metadata was deemed sufficiently complex to be considered separate from the other descriptive metadata.

The consequences of using the <mods:extension> element to refer to the degree level and discipline are unclear. Related to this issue, the ETD community would benefit from a controlled vocabulary of disciplines.

Further work could be done on standardizing the information in the <mods:recordInfo> wrapper element, especially for identifying the source of the content and establishing a standard identifier for MODS records. For MARC records, much of this is work is coordinated by OCLC, but the ETD environment is decentralized and standardization may be more difficult.

Theoretically, MODS has the capability for describing compound objects through the <mods:relatedItem> element, but it is unclear how this should be done in practice. The problems include standardizing descriptive elements, developing systems that create metadata in an efficient manner, and managing this complex information in systems for search, retrieval, and display.

TDL is creating MODS records though crosswalks from MARC and Dublin Core records, but these crosswalks have limitations. MARC records typically do not have all of the information specified in our application profile and Dublin Core records often lack sufficient descriptive specificity. Ideally, ETD management systems will be developed that create MODS records natively during the ETD submission process. If other types of records are desired, namely MARC or DC records, they could be derived from the MODS using standardized crosswalks.

5. MANAKIN AND MODS

Manakin, developed by Texas A&M University, is a software package that allows institutions to customize the DSpace user interface (Phillips, et al., 2005). The Manakin project has five goals: 1) Allow each community to maintain a distinct look and feel, 2) Separate business logic from stylistic design, 3) Establish an interface-level component architecture, 4) Enable internationalisation and localization of content, and 5) Provide an alternative interface that does not replace the existing JSP interface. TDL, the first implementer, is using Manakin to extend a customized representation of the ETD collection into the general TDL website.

Given highly specific metadata encoded in MODS and the ability to customize the user interface, TDL has a great deal of flexibility in designing a portal to its ETD collection. The TDL ETD collection conforms to the general design of the overall TDL site and has features such as a pre-formatted citation for each ETD.

6. CONCLUSION

TDL adopted MODS as a common metadata standard for ETDs because of its benefits over MARC and Dublin Core. TDL implemented MODS in a union catalog that includes ETDs from two institutions: the University of Texas and Texas A&M University. Manakin complements this rich ETD metadata by allowing the development of customized user interfaces for DSpace collections.

Although MODS represents a step forward, challenges remain, including publisher metadata, rights metadata, the use of non-MODS extensions, standardization of record information, the description of compound objects, and the development of ETD systems that use MODS natively. TDL will continue to investigate these issues as well as promoting the development of ETD systems in Texas universities.

7. TABLE COMPARING ETD-MS AND TDL-MODS ELEMENTS

Table 1. Comparison of ETD-MS and MODS

Descriptive element	M/O	ETD-MS	TDL MODS
Title information	M	dc.title dc.title.alternative	titleInfo title subTitle
Name of author	M	dc.creator	name namePart role roleTerm
Subjects	M	dc.subject dc.coverage	subject topic geographic temporal
Abstract	M	dc.description dc.description.note dc.description.release	abstract
Publisher	O	dc.publisher	Not specified
Name of thesis advisor	M	dc.contributor dc.contributor.role	name namePart role roleTerm
Name of committee members	O	dc.contributor	name namePart role roleTerm
Origin Information	M	dc.date	originInfo dateCreated dateIssued
Genre	M	dc.type	genre
Physical description	M	dc.format	physicalDescription form internetMediaType digitalOrigin
Identifier	M	dc.identifier	identifier
Language	M	dc.language	language languageTerm
Rights	O	dc.rights	Not specified
Degree Information	M	thesis.degree.level thesis.degree.discipline	extension etd:degree etd:name etd:level etd:discipline
Name of degree grantor	M	thesis.degree.grantor	name namePart role roleTerm
Type of resource	M	Not specified	typeOfResource
Location	M	Not specified	location url
Record Information	M	Not specified	recordInfo recordContentSource

			recordCreationDate recordChangeDate recordIdentifier

8. REFERENCES

Atkins, A., E. Fox, R. French, and H. Suleman. (2006). *ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations, version 1.00, revision 2.* NDLTD. Retrieved June 6, 2006 from http://www.ndltd.org/standards/metadata/current.html.

Frodl, C. and N. Korb. (2005). XMetaDiss meets ETD-MS. *ETD2005: The 8th International Symposium on Electronic Theses and Dissertations*. Retrieved June 6, 2006 from http://adt.caul.edu.au/etd2005/papers/051Frodl.pdf.

Guenther, R. S. (2003). MODS: The Metadata Object Description Schema. *portal: Libraries and the Academy*, 3, no. 1, 137-150.

Hoover, L. and R. E. Woolverton, Jr. (2003). Cataloging and Treatment of Theses, Dissertations, and ETDs. *Technical Services Quarterly*, 20, no. 4, 3-57.

Joint Steering Committee. (2005). *Anglo-American Cataloguing Rules, Second Edition, 2002 Revision*. Chicago, IL: ALA, CLA, CILIP.

Library of Congress. (2005). *MODS User Guidelines, Version 3.* Washington, DC: Library of Congress. Retrieved June 6, 2006 from http://www.loc.gov/standards/mods/v3/mods-userguide.html.

Metadata Working Group, Texas Digital Library. (2005). *MODS Application Profile for Electronic Theses and Dissertations.* Austin, TX: Texas Digital Library. Retrieved June 6, 2006 from http://www.tdl.org/projects/metadata/tdlappprofile.pdf.

OCLC. (2003). *Bibliographic Formats and Standards, Third Edition.* Dublin, OH: OCLC. Retrieved June 6, 2006 from http://www.oclc.org/bibformats/en/.

Phillips, S., C. Green, J. Leggett, A. Maslov, A. Mikeal, and B. Surratt. (2005). *Manakin Developer's Guide: The Second Version of the DSpace XML UI Project.* College Station, TX: Digital Initiatives, Research, and Technology, Texas A&M University Library. Retrieved June 6, 2006 from http://di.tamu.edu/projects/xmlui/manakin/resources/DevelopersGuide.pdf.