7-1-2009

Defining Best Practices in Electronic Thesis and Dissertation Metadata

Rebecca L. Lubas
Claremont University Consortium

Journal of Library Metadata

Defining Best Practices in Electronic Thesis and Dissertation Metadata

Rebecca L. Lubas[a]

[a] Cataloging and Discovery Services at the University of New Mexico Libraries, University of New Mexico, Albuquerque, New Mexico, USA

Online publication date: 10 December 2009

PLEASE SCROLL DOWN FOR ARTICLE

Defining Best Practices in Electronic Thesis and Dissertation Metadata

REBECCA L. LUBAS

*Cataloging and Discovery Services at the University of New Mexico Libraries,
University of New Mexico, Albuquerque, New Mexico, USA*

The University of New Mexico will mandate in 2009 that theses and dissertations be submitted in electronic form as the copy of record. These documents will reside in the university's digital repository, operated on a DSpace platform. This article reviews practices for thesis and dissertation metadata creation with a focus on DSpace instances, best practice recommendations for author-submitted metadata, recommendations for subject analysis, and training for metadata practitioners. The article recommends processes for author submission, metadata quality control and enhancement, and crosswalking of the metadata to the library's catalog to maximize discovery.

KEYWORDS University of New Mexico, UNM, electronic theses and dissertations, ETDs, Dublin Core, author-generated metadata, metadata best practices, authority control, MARC, DSpace

INTRODUCTION

As is the case for most university libraries, the University of New Mexico (UNM) has collected, preserved, and served theses and dissertations using the traditional library metadata combination of the Anglo-American Cataloging Rules, 2nd Edition (AACR2), Machine-Readable Cataloging (MARC), its integrated library system (ILS), and an indexing and abstracting service. Users can currently expect to find the UNM Libraries' collection of theses and dissertations represented in the university's online catalog, OCLC's WorldCat, and in Dissertations Abstracts. Starting in the summer semester 2009, UNM

Address correspondence to Rebecca L. Lubas, Director of Cataloging and Discovery Services at the University of New Mexico Libraries, 1 University of New Mexico, Albuquerque, NM 87131, USA. E-mail: rlubas@unm.edu

will require theses and dissertation authors to submit their work to the university's institutional repository, DSpaceUNM. Paper versions will no longer be collected. There are a number of concerns that arise from the "electronic only" policy and there are key concerns to address with regard to the process. Many of these concerns can be addressed with sound metadata decisions and practices.

This study seeks to discover the current best practices in electronic theses and dissertation (ETD) deposit, methods for author-submitted metadata, methods for enhancing that metadata, and the skills required of catalogers/metadata librarians to shepherd the process. This study will also examine the ETDs and associated metadata collected in the UNM ETD pilot stage to aid in determining what enhancements should be added in both access and metadata.

QUESTIONS

A major question to consider is the following: Once theses and dissertations are submitted to DSpaceUNM, should the titles be represented in the libraries' catalog? If the work resides in this repository alone, information about the collection would be divided between two silos, creating a barrier to discovery. As of this writing, UNM is only beginning to implement a federated search that would encompass multiple databases with a single search. Hence, it may be best at this stage of search-tool development to deposit metadata in both the institutional repository and in the catalog. With double deposit of metadata required, finding the most efficient way to prepare the author-submitted metadata would be of benefit. Portability of the metadata must also be considered, for harvesting of the documents and for the inevitable future migration to systems yet unimagined.

Another workflow concern involves the enriching of author-submitted metadata. Theses and dissertation authors provide basic metadata such as name, abstract, key words, and department name as part of the submission process. In some ways, this is no different than in the print world, in which the information is taken from the author-provided thesis title page. However, in traditional library practice, that information is transformed into a surrogate of the work using well-developed standards and controlled vocabularies. A new format for theses and dissertations requires that the library (or other caretakers of the collection) determine which controls should be applied by a cataloger/metadata specialist during the submission and publication process, and at which points they should be applied.

Since creation of electronic theses and dissertation metadata will require that additional standards be employed in the library, skill acquisition of the staff is another concern. As with many research libraries, UNM has a staff of experienced catalogers, practiced in the creation of AACR2/MARC

metadata. UNM catalogers have also been trained in the standards of the Name Authority Program of the Program for Cooperative Cataloging (NACO). Transferring these skills to the Dublin Core metadata enhancement process and crosswalking from DC to MARC will require training and practice.

Authority control of names is one of the most neglected areas of metadata creation outside traditional library workflows. This study will consider the value of identifying and using a standard form of the author's name as part of the metadata enhancement and quality control process.

PILOT ETDS AT UNM

DSpaceUNM, much like other instances of this digital repository platform, is used as a digital archive for the institution's research and creative works. UNM established DSpace as its institutional repository based on the provost's decision that there will be only one instance for the whole university, which includes the main campus, medical, and law campuses, and the four branch campuses in other cities. UNM Libraries implemented DSpace in 2005.

Prior to mandatory electronic submission of theses and dissertations, some students voluntarily added electronic versions of their work to DSpace-UNM, while still being required to submit a paper copy. UNM's Office of Graduate Studies serves as the approver for the electronic submissions. As of March 2009, there were 31 dissertations and 42 theses in the repository.

The paper copies continued to be bound and cataloged in OCLC and LIBROS (the Libraries' Innovative Millennium-based ILS). Prior to this study in spring 2009, the author-submitted Dublin Core metadata was not reviewed in detail or enhanced by a cataloger or metadata specialist. There was no connection between the metadata for the electronic version and paper version; no link for the electronic version was added to the MARC metadata for the paper version.

In fall 2008, with the target date for electronic only submission on the horizon, UNM Libraries personnel in the Center for Southwest Research (UNM's archive and special collections), Library Information and Technology, and Cataloging and Discovery Services began exploring how best to carry the goal of making the theses and dissertations discoverable and accessible in the hybrid world of the post-paper era. As there are no current resource allocations available for digitizing the paper collection prior to mandatory submission, the need for the ability to search the whole collection is also a consideration. For example, it would require two searches (and the knowledge that one needs to search two places) to recall all the theses by a given advisor. UNM staff looked toward other libraries that had already implemented ETD programs for clues to prepare for a more robust ETD service and collection.

REVIEW OF LITERATURE AND CURRENT BEST PRACTICES

Theses and dissertations were an early target for electronic archiving and distribution. The year 2009 marks the twelfth conference devoted solely to the subject.[1] ETDs present many logistical issues. Submission, authentication, distribution, and preservation are major processes requiring careful planning to maintain the integrity of these products of a university's intellectual output. Metadata creation is but one aspect. The third version of the digital-scholarship.org's ETD bibliography, begun in 2005, includes only two articles focused solely on ETD metadata practice (Bailey, 2009).

This number does not exhaust the available guidance and opinion for ETD metadata. Many overall guides for implementing ETDs address the topic of metadata. At the 2004 ETD conference, five presentations covered metadata to some extent, with practices presented from repositories in North America and in Europe. Many institutions implementing ETDs employed the qualified Dublin Core fields crafted for theses and dissertations by the University of Edinburgh (Jones, 2004). Administrative metadata is usually created at point of submission, much of it machine generated. Rights metadata is often set as a matter of policy up front, and automatically added to the process. Descriptive metadata can be deceptively simple. In a repository such as DSpace, authors can easily submit basic elements of descriptive metadata, and their input is contained in metadata standard such as qualified Dublin Core. ETDs are full-text searchable in DSpace and other repository systems, so the need for a metadata quality-control process or application of a controlled vocabulary may not appear paramount.

Common threads appearing throughout the literature are the inconsistency of author-generated metadata and the need for quality control, the time required for expert metadata enhancement, and the limits of the Dublin Core element set.

In 2004, Janick and McLaughlin presented at an overview of the ETD program at Drexel University, which uses DSpace. They included a critique of the descriptive metadata available in DSpace, citing its employ of the minimalist Dublin Core as a weakness. Specific metadata elements lacking include date degree is awarded, type of degree, advisors and committee members, date of defense, and contact information for the author. They also list some metadata labels that are available as being of little value to an ETD such as alternate title, series or report title, sponsors, and additional authors. Some of these elements have potential for ETD metadata, for example, "sponsor" could be useful if a dissertation was completed with the aid of research grants.

At the time of the presentation, Drexel had a cataloger enhance the data with Library of Congress subject headings after submission. Jancik and McLaughlin cite a discovery need, criticizing DSpace for not making it easy

to bring together a list of all the theses and dissertations completed by a certain department.

A follow-up call to Drexel in February 2009 confirmed that they continued to have a cataloger enhance the subject keywords in the DSpace metadata with Library of Congress subject headings.[2] They continue to separately catalog the paper copy in MARC in their local catalog and in OCLC, and the MARC record includes the 856 tag to link to the DSpace electronic version. The MARC metadata does not necessarily include all the elements in the Dublin Core or vice versa. For example, the abstract is required in DSpace but not necessarily present in the MARC version.

Also in 2004, El-Sherbini and Klim documented the metadata creation process for the then-emerging OhioLINK ETD Center, an online center for members of this large library consortium. At that time, the emphasis remained on creation of MARC records in OCLC and the OhioLINK catalog. Catalogers enhanced the MARC by obtaining abstracts from the author-submitted metadata. If, as there often were during a transition period, both library-collected paper and electronic versions, the record for the paper version was enhanced and linked to the electronic version.

The publically accessible author submission form in the OhioLINK ETD Center (viewed in March 2009) accepts a number of author-generated elements. In addition to the traditional name elements, it gives the author the option of providing e-mail contact information and making it publically available. The form includes the traditional title, abstract, and key word elements, and provides a drop-down menu of the ProQuest UMI vocabulary of subject headings. These headings are broad-level, such as Biology or Library Science. Drop-down boxes also provide lists of degree names and departments. Names of advisors and committee members can be added, with drop-down menus to identify the roles of the individuals named. Further screens give authors the option of choosing between copyright statements, including a creative commons license wording. One can view the metadata in MARC, basic Dublin Core (DC), the DC-based Electronic Theses and Dissertation Metadata Standard (ETD-MS), and html.

An OhioLINK member library, Kent State, has used the ability to harvest metadata from the ETD Center to achieve instantaneous discovery in its local catalog (McCutcheon, Kreyche, Maurer, & Nickerson, 2008). Kent State harvests metadata from the Center using a Perl script with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The result is immediate population of MARC metadata in the local catalog with a link to the full text. This machine generated metadata is then enhanced by a cataloger who then contributes full-level MARC records to OCLC. Therefore, Library of Congress subjects are added, punctuation is standardized, and AACR2-prescribed notes are created. Kent State uses author-generated metadata as a base for full level MARC cataloging, giving them the advantage of immediate discovery and also the richness of hand-built MARC.

By 2006, alternatives to Dublin Core were actively in use. The Texas Digital Library chose the Metadata Object Description Schema (MODS), citing the common criticism that DC had "not many elements" (Surratt, 2006). Despite this pervasive complaint, unqualified DC continues to proliferate via OAI-PMH (Jordan & Shearer, 2006).

Kansas State University decided to end paper archiving of theses and dissertations altogether and preserve and present the electronic form of the content in a DSpace instance.[3] They quickly concluded that better author-generated metadata would improve the final product and took steps to improve the chances of doing so at the beginning of the process by enhancing the submission form. Their form employs numerous drop-down menus giving the authors official forms of the names of degrees, departments, and names of professors. They provide instructions online and have found thus far that students rarely have questions about the submission process. Library personnel perform a quality check on the data by opening the text (a pdf file) and checking against the metadata. KSU provides additional access to their ETDs by running the DSpace Dublin Core elements through a MARCit style sheet to create metadata usable in the library catalog. There is constant data added for consistency with the local catalog. Another quality check is performed, mainly to look for elements that could cause retrieval problems in the MARC-based catalog such as improperly encoding of titles with initial articles (which are skipped in MARC searching). KSU reports that since enhancing the input form, personal and department name data is of very high quality, needing little intervention. Finally, the MARC records are uploaded into OCLC for WorldCat and exported into the local catalog.

There are considerations beyond the institution's local repository and the local catalog. As early as 2002, harvesting for cross-institutional electronic theses and dissertations search was being discussed at the ETD conference. Hussein Suleman presented a practical guide to creating an open archive of ETDs via OAI-harvesting based on work at Virginia Polytechnic Institute and State University. He defined the lack of interoperability of different metadata schema as a barrier to cross searching and harvesting, and recommended using XML-coded unqualified Dublin Core as the common metadata language. Despite much criticism for being oversimplified, half a decade later Dublin Core has emerged as a practical approach to metadata language switching. Today, the Networked Digital Library of Theses and Dissertations can search ETDs from more than 90 member institutions.[4]

While during the early days the use of a simplified metadata element set such as Dublin Core may have seemed limiting, over the course of a decade of experience with electronic theses and dissertations metadata reveals that blending the use of qualified Dublin Core with harvesting and crosswalks, plus creating tools to encourage better results from author-generated metadata have proved useful.

LOOKING AT UNM'S VOLUNTEER ETD METADATA

Currently, UNM Libraries collects, binds, and creates metadata in World-Cat and in the local catalog for paper theses and dissertations. Electronic submission has been encouraged but not mandatory. The UNM cataloging department creates full-level AACR2/MARC records using consistent forms of the department names, Library of Congress subject headings, and Library of Congress call numbers. Abstracts and author-supplied key words are not used.

As of March 2009, there were 73 ETDs in DSpaceUNM, organized in DSpace collections by department (and in some cases, degree program) name. There is no connection between the electronic version and the paper version, no link is made in the MARC record to DSpace.

Currently, DSpaceUNM uses a modified submission form to capture some the unique aspects of ETDs for description in modified Dublin Core. The form prompts the submitter to give the metadata, represented in Table 1.

A notable benefit of the DSpace version of the ETDs is their immediate availability after approval. The text of the accepted ETDs are available to the searching public while the paper copies may still wait at the bindery and/or be in a queue for metadata creation for the local catalog.

A review of the metadata from this 73-document sample revealed that they contain potential helpful metadata for the MARC versions. The abstracts, not only provide the searcher with a better scope of the work, but would also enrich key-word searching in the local catalog. The author-supplied key-words in the dc.subject space, not only provide more keywords for searching, but can also help the cataloger performing quality control and enrichment with more clues for subject analysis. Since theses and dissertations are

TABLE 1 UNM ETD Metadata Input Guide

Form Name	DC Label	Input Method
Author	dc.contributor.author	Free text
Title	dc.title	Free text
Date of Graduation	dc.date.submitted	Menu with choice of semester/year
Publication Type	dc.type	Menu (Thesis, Dissertation, or Report)
Degree Name	dc.description.degree	Free text
Degree Level	dc.description.level	Menu (Doctoral, Master's)
Department	dc.description.department	Free text
Major Professor	dc.description.advisor	Free text
Committee Members	dc.description.committee-member	Free text
Language	dc.language	Menu
Subject Keywords	dc.subject	Free text
Abstract	dc.description.abstract	Free text
Sponsors	dc.description.sponsorship	Free text

theoretically new research, they are often challenging to analyze with controlled vocabularies. Each additional key word can enhance the discoverability of the work. Most of the 73 submitters took the opportunity to add their own key words, only one did not.

The addition of committee members, not previously included in UNM's AACR2/MARC metadata, could be of additional benefit if crosswalked to the library's catalog, enabling searchers to find research products of professors not necessarily the primary advisor.

The author-submitted department names proved problematic. In this free-text field, submitters came up with varying forms of department names, ranging from using ampersands instead of the word *and* to omitting key words in the name, for example, "computer engineering" instead of "electrical and computer engineering." In some cases, this led to confusion about which collection to submit the work to. Five theses left out the department name entirely (although the thesis could still be identified by the collection to which it was submitted).

Fifteen theses and dissertations were submitted with titles and author names in all capital letters. While this is not a problem for free-text–keyword searching, it is a problem for repurposing the data in other metadata standards.

RECOMMENDATIONS TOWARD BETTER PRACTICES

Two Metadata Records are Better than One

Library collections are still best represented by their MARC metadata. This will remain true for some time in the future, as research library ILSs can average 2 million or more titles in their holdings represented by MARC. There is no more complete venue for MARC metadata than OCLC. Representing a library's collections in OCLC's WorldCat is a powerful way to reach searchers, especially as OCLC is making WorldCat.org available in an increasing number of venues, such as Facebook and smart phone applications. With an increasing number of libraries selecting WorldCat Local for their discovery layer, there is yet another incentive to deposit ETD metadata in OCLC.

There is, however, incentive for creating the ETD metadata in Dublin Core. Major institutional repository software is Dublin Core based. Dublin Core is also harvester friendly.

While OCLC is equipped to ingest Dublin Core data directly, enhancement to full, or at least fuller, levels in OCLC adds value. Many of the institutions that have implemented ETDs have used crosswalks to bring the Dublin Core metadata into their library's main catalog. The point of interaction with OCLC provides an opportunity for the library to perform authority work, the normalization and disambiguation of author names. As a thesis or dissertation is often a writer's first work, this is an opportunity to get direct

input from the author for their preferred form of name, and to distinguish the name amongst common names. It is much easier to obtain this information when there is an institutional connection to the author than later in the publication life cycle.

Committing to a Dublin Core–based record in the institutional repository and then crosswalking to MARC for upload to OCLC and the local ILS provides a good range of exposure of the metadata to search. The metadata is open for harvesting, traditional library catalog searching, and the worldwide users of OCLC services.

Submission to an institutional repository increases discovery earlier in the process. An ETD can be available immediately after acceptance and is full-text searchable in DSpace and other institutional repositories. The metadata enhancement can be performed later, not holding up accessibility as it does for paper copies awaiting binding and metadata creation.

Improving Author Generated Metadata

Based on the practices of the libraries described in the review and a scrutiny of the volunteer-submitted metadata in DSpaceUNM, encouraging higher-quality, author-generated metadata at the beginning of the process would be beneficial both for increasing discovery of the ETDs within the repository as well as for repurposing the metadata for other searches.

As Kansas State's experience demonstrates, the use of menus of department names and professor's names ensures consistency and reduces the amount of regularization of the names in the quality-control process. Using authorized forms of department names from the beginning of the process reduces authority work if the records are to be uploaded to WorldCat in full-level MARC. In DSpace in particular, this can be accomplished by regularizing department names via the collection/community organization as well as by listing in the metadata entry drop-down menus rather than as free-text entry. Established theses advisors can also be provided in a drop-down list. These pre-enhancement controls improve discovery by providing consistent metadata for searches early in the life of the digital document. For institutions using DSpace and other widely used platforms, the content is also discoverable via Google, Google Scholar, and Scientific Commons.

Making training available to thesis and dissertation authors should have benefits as well. While it is unrealistic to expect a single-submission author to produce perfect ISBD punctuation or use of subject vocabularies, for example, training can reduce/eliminate some of the more basic problems encountered in the UNM sample batch, such as the input of names and titles in all capital letters. It is also an opportunity to encourage the students to give more voluntarily information, such as key words and additional name information to aid in later subject analysis and authority work.

Enhancing the Dublin Core Metadata

Many of the libraries that input ETDs in their repositories and catalog them in WorldCat and the local catalog tend to focus on manual enrichment of the MARC by a cataloger (whether generated from the Dublin Core or not). Despite the full-text search in DSpace, enhancing the Dublin Core metadata directly in DSpace before the MARC conversion adds value to the search. AACR2/MARC-trained catalogers can apply their skills to working directly with Dublin Core in DSpace. Library of Congress subject headings have a home in qualified Dublin Core (dc.subject.lcsh).

It is recommended that the author-generated metadata only be enhanced, not deleted, except for clear errors such as misspellings. All author-generated key words should remain. These enhancements will be carried over to the MARC if one is using the DC as a base record for the MARC via crosswalk.

Training

Despite is simplicity, Dublin Core requires some familiarly and training to achieve good results. One of the most common causes for confusion in the use of Dublin Core is ambiguity and vagueness of the definitions of the elements (Park & Childress, 2009). Training in the institution's specific application of Dublin Core can address this issue. In the case of ETDs, with specific needs for sponsoring departments and advisors to appear in the Contributor elements, as well as the challenges of subject analysis, specific training focused on ETD Dublin Core metadata creation, even if the cataloger is experienced in Dublin Core, is recommended.

A key skill required, though less tangible than training in a specific standard, is flexibility and awareness of trends (Park & Lu, 2009). Creating the best ETD metadata for user discovery will require awareness of the latest developments in user behaviors and the most current search tools and the harvesting methods behind them.

UNM's ETD Workflow

Before the summer 2009 semester, UNM libraries, in cooperation with the Office of Graduate Studies, edited the department names in DSpaceUNM to be consistent with the list of department names that had been used as authoritative for the paper theses and dissertations collections. The submission instructions for authors were also edited to more strongly encourage authors to give key words. UNM information technology and cataloging personnel also tested a Dublin Core-to-MARC crosswalk in preparation.

After one-on-one training with a cataloger experienced in Dublin Core metadata creation, UNM catalogers will enhance the Dublin Core metadata in DSpaceUNM, adding subject headings and editing the free-text fields. After the crosswalk is applied, the resulting metadata will be examined by a cataloger a second time for MARC coding quality. Then the metadata will be uploaded to OCLC and UNM's LIBROS.

CONCLUSION AND LOOKING FORWARD

As the libraries continue to strive toward stronger search solutions and more efficient metadata creation, combining author-generated metadata with quality control procedures, metadata enhancement, and exposure of the content to multiple venues is a good transitional step to achieve discovery.

There are outstanding, unresolved questions still to be answered about the long-term stability of institutional repositories. Research on preservation of ETDs is still needed. At the time of the writing of this article, it is listed as a major topic of the 2009 ETD Conference. ETDs are now scattered in OCLC, library catalogs, and search-engine tools such as Google Scholar. Where are searchers most often turning to search for theses and dissertations? Are they searching library catalogs for known-item searches and harvesters and search engines for general discovery? What searches lead to the most satisfactory outcome for the searcher—advisor names? keywords? controlled vocabularies?

ETDs exist in a hybrid world. The caretakers of these collections need to remain consistent yet flexible, and increase awareness of user behavior.

ACKNOWLEDGMENTS

NOTES

1. ETD 2009 conference http://www.library.pitt.edu/etd2009/program.html

2. Conversation with Robert Sieczkiewicz, University Archivist, Drexel University on February 18, 2009.

3. Conference call with Martin Courtois and Michelle Turvey-Welch at Kansas State University on February 13, 2009. ETD instructions for students are online at http://www.k-state.edu/grad/etdr/submit/subkrex.htm

4. NDLTD home page, viewed March 19, 2009, http://www.ndltd.org/

REFERENCES

Bailey, C. (2009). Electronic theses and dissertations bibliography, version 3. Retrieved March 26, 2009, from http://www.digital-scholarship.org/etdb/etdb.htm

El-Sherbini, M., & Klim, G. (2004). Metadata and cataloging practices. *The Electronic Library*, *22*(3), 238–248.

Janick, S., & McLaughlin, T. (2004). Using DSpace to administer an ETD program: The Drexel experience. Retrieved March 26, 2009, from http://www.uky.edu/ETD/ETD2004/janick/etd062004.ppt

Jones, R. (2004). Using DSpace for e-theses at Edinburgh University Library. Retrieved March 26, 2009, from http://www.thesesalive.ac.uk/archive/DSpaceUGMeeting.ppt

Jordan, M., & Shearer, K. (2006). A metadata application profile for Canadian repositories. Retrieved March 26, 2009, from http://www6.bibl.ulaval.ca:8080/etd2006/pages/papers/SP6_Mark_Jordan.ppt

McCutcheon, S., Kreyche, M., Maurer, M., & Nickerson, J. (2008). Morphing metadata: Maximizing access to electronic theses and dissertations. *Library Hi Tech*, *26*(1), 41–57.

Park, J. & Lu, C. (2009). Metadata professionals: Roles and competencies as reflected in job announcements, 2003–2006. *Cataloging & Classification Quarterly*, *47*, 145–160.

Park, J., & Childress, E. (2009). Dublin Core metadata semantics: An analysis of the perspective of information professionals. *Journal of Information Science*, *20*(10), 1–15.

Surratt, B. (2006). MODS meets Manakin: Innovations in the Texas Digital Library's thesis and dissertation collection. Retrieved March 26, 2009, from http://www6.bibl.ulaval.ca:8080/etd2006/pages/papers/SP6_Brian_SURRATT.pdf