Cataloging & Classification Quarterly, 47:297–308, 2009

ISSN: 0163-9374 print / 1544-4554 online
DOI: 10.1080/01639370902737323

Routledge
Taylor & Francis Group

Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries

MICHAEL BOOCK and SUE KUNDA
Oregon State University Libraries, Corvallis, Oregon, USA

In July 2005, the Oregon State University Libraries began accepting electronic versions of student theses and dissertations into ScholarsArchive@OSU, the library's institutional repository. By January 2007, all Oregon State University graduate students were required to deposit their final research. This article compares past processes and workflows for print theses and dissertations with the present workflow for electronic. We provide the rationale for changes and review the cost- and time-savings produced. We describe the changing roles of students, technicians, and librarians in the metadata process as well as the value of students describing their own work.

KEYWORDS ETDs, metadata, workflow, Oregon State University

INTRODUCTION

In July 2005, the Oregon State University (OSU) Libraries began accepting electronic versions of student theses and dissertations via ScholarsArchive@OSU, the library's institutional repository (IR). By January 2007, all OSU graduate students were required to deposit their final research to the Electronic Theses and Dissertations (ETD) Collection in ScholarsArchive@OSU. OSU is one of sixteen members of the Networked Digital Library of Theses and Dissertations association to require deposit of all ETDs.[1]

Using the DSpace digital repository platform, the Oregon State University Graduate School and library staff coordinated efforts to provide a seamless submission process as well as a self-explanatory template for

Received July 2008; revised September 2008; accepted October 2008.

Address correspondence to Sue Kunda, MLS, Digital Production Librarian, Oregon State University Libraries, 121 The Valley Library, Corvallis, OR 97331. E-mail: sue.kunda@oregonstate.edu

resource description. Students create and attach basic descriptive metadata for their research; the Graduate School validates the thesis or dissertation; library staff review the student-submitted metadata, ensure that name headings are authorized, and add Library of Congress Subject Headings (LCSH). DSpace automatically generates structural and administrative metadata. The library uses a modified version of MarcEdit, a homegrown cataloging utility, to map the DSpace Dublin Core metadata to MARC. During this process, the utility imports the metadata from the IR into the library online catalog and WorldCat.

This article compares past processes and workflows for print T/Ds with the present workflow for ETDs. We provide the rationale for changes and review the cost- and time-savings produced by the revised workflow. We describe the metadata mapping utility that generates MARC records in the local online catalog and WorldCat from Dublin Core repository records. We describe the responsibilities of the library and graduate school in the overall process and the specific responsibilities of students, library technicians, and catalog librarians in the assignment of metadata and processing of ETDs. We also discuss future considerations including the possibility of no longer requiring a print copy to be submitted and no longer supplying LCSH.

BACKGROUND

Like many academic institutions, OSU has been interested in electronically capturing student T/Ds ever since they were first produced using word processors. In January 1997, Virginia Tech University, under the direction of Ed Fox, began requiring submission of theses and dissertations to a digital repository.[2] West Virginia followed in 1998 with twelve other institutions worldwide (nine in the United States and Germany, Brazil, and South Africa) adding the requirement since.[3]

Oregon State University Libraries began discussing the possibility of capturing, storing, and providing access to ETDs with the Graduate School in 1998. Although there was no agreement to move forward at that time, a good relationship was formed. A lack of trusted software and a lack of clarity regarding what each department would be responsible for in terms of staffing and funding of technical support were early impediments. There was general agreement that these issues were not insurmountable and that the two could come to an agreement.

In 2004, the Libraries began investigating IRs and agreed that ETDs would be a natural fit for an IR at OSU. Reasons include:

- T/Ds constitute and represent the research conducted by the university, and the Libraries are required to archive them in perpetuity.[4]

- Unlike faculty, students could be required to deposit their research, resulting in guaranteed content growth for the IR.
- The Libraries and Graduate School shared a philosophy that OSU's student research should be more widely accessible.

After investigating repository options, including Fedora and EPrints, DSpace was selected for the university's IR software. The software is relatively easy to install and maintain, it is customizable to meet unique needs, and it offers coherent workflow capabilities, authorization management, and a straightforward submission process. The Libraries were also interested in working with an open source community with a committed user base.[5]

The Libraries and Graduate School began meeting once again and, shortly after DSpace was installed, the two agreed to proceed with a pilot program. A single department that had previously expressed an interest in ETDs was selected; they began requiring their students to submit their T/Ds electronically in August 2005. The Graduate School communicated to the academic deans progress and plans for requiring additional departments to submit T/Ds. Quickly, many other departments expressed an interest in participating. There were so few problems with the workflow and process that all agreed to begin requiring PhD dissertations to be submitted starting in July 2006 and masters theses in January 2007.

WORKFLOWS

Print

From 1902 to July 2005, OSU Libraries was responsible for storing two copies of all print T/Ds produced at the University. One copy was held in the circulating collection and a second was stored as an archival, backup copy at the library's offsite storage facility. The library's workflow for handling print-only T/Ds is probably typical (see Table 1). We present it here to contrast it with our workflow for ETDs and to demonstrate the cost and time savings of moving from paper to electronic.

Once approved by the Graduate School, a library binding technician retrieved two copies of each print T/D from the Graduate School and generated a list for the cataloging department. In addition, the binding technician inspected each copy for incorrect paging, missing pages, pages with inadequate gutter margins, proper paper quality, and appropriate signatures. Both copies of theses were collated and sent to a commercial binder. One dissertation copy went to the bindery and one went to UMI for filming. After returning from UMI, the second copy of the dissertation was also sent to the binder.

Using the list of theses generated by the binding technician, a cataloging technician created LC call numbers, essentially just adding a cutter for author because all print T/Ds were shelved in a single LC class number. The

TABLE 1 Print and ETD Comparison

Task	Print T/D (minutes/100 items)	ETD (minutes/100 items)
Retrieve theses from Graduate School	120	60
Generate list of theses	60	0
Examine and collate theses	2000	1000
Prepare/send theses to bindery	60	0
Prepare/send dissertations to UMI	60	30
Prepare theses for storage	60	60
Create call numbers from list of theses	120	0
Create brief bib record with call number	480	0
Create full MARC record (physical description/local subject heading)	1080	0
Review student-created metadata; authorize names; add descriptive notes	0	500
Assign LCSH	3250	3250
Download and overlay record in OPAC	100	0
Map/review/export DSpace metadata to OPAC and WorldCat	0	950
Receive UMI dissertations; prepare/send dissertations to bindery	60	0
Receive/examine theses/dissertations upon return from bindery	300	0
Shelve bound theses/dissertations in stacks	60	0
Shelve storage copies of theses/dissertations	60	60
Total time spent (minutes/100 items)	7870	6020
Total time spent (minutes/item)	~80	~60
Total processing cost (per item)	$23.23	$18.87
Total processing cost (annual)	$11,615.00	$9,435.00

binding technician then created a brief bibliographic record, including this call number, and attached an item record with barcode. After inserting a copy of the bib record into the T/D, the T/Ds was again sent to a cataloging technician who created a more complete MARC record. This record included physical description and publication information as well as a local subject heading based on the department granting the degree.

Until 2003, catalog librarians assigned LCSH and library technicians entered the catalogers' handwritten records into WorldCat. After losing two catalog librarians within several months of each other, the library decided to train higher-level technicians to do subject analysis, a practice that continues today. After assigning LCSH, the library technician downloaded the completed bib record to the library's online public access catalog. Student workers shelved the circulation copy when it returned from the binder while a circulation technician transported the second, archival copy to the Libraries' storage facility. In total, each T/D was handled by five different people and changed hands five times from start to finish. The total processing time for each T/D was approximately 80 minutes and cost the library $23.23.

Electronic

ScholarsArchive@OSU, Oregon State University's digital service for gathering, representing, and recording the work of its research and teaching community, operates on the DSpace digital repository platform and hosts the university's ETD collection. The Graduate School and Libraries coordinated efforts to provide a seamless and easily understood ETD submission process.[6] Graduate students log in using their university-assigned Lightweight Directory Access Protocol (LDAP) login and password and are automatically authorized to submit to the Electronic Theses and Dissertations collection in ScholarsArchive@OSU.

The Libraries capture 24 descriptive and administrative metadata elements (see Appendix), most of which are recommended in the Networked Digital Library of Theses and Dissertations (NDLTD) ETD metadata standard (ETD-MS). The ETD-MS provides "a standard set of metadata elements used to describe an electronic thesis or dissertation" and guidelines for their use in various environments.[7] The DSpace digital repository software uses a qualified version of the Dublin Core metadata schema based on the Dublin Core Libraries Working Group Application Profile (LAP),[8] and automatically generates administrative metadata including provenance (who submitted the items and at what date and time), rights management (who has access to the T/D), source (where the file originated), and technical information (number of bytes and file management data).

Compare the print T/D workflow with the ETD workflow detailed in Table 1. Much of the descriptive cataloging work once done by cataloging technicians is now completed by the graduate student as she or he works through the submission process. The descriptive metadata that students enter includes:

- author
- title
- advisor
- committee member
- abstract
- keywords
- degree name
- degree level
- college in which they are receiving their degree
- granting institution
- academic department
- graduation date
- language

Students have the option of applying a Creative Commons license during submission. They grant a non-exclusive distribution license to Oregon State

University before finally uploading a PDF version of their T/D. If the student has accompanying materials to upload, such as maps or data, that too is uploaded.

Upon successful submission, the Graduate School receives notification that an ETD was submitted. A thesis editor with the Graduate School validates the ETD as well as a single print copy that is still required. Upon validation, a notification is sent to two library technicians who share the work of reviewing the student-submitted metadata, making corrections as appropriate. Corrections often include ensuring information provided during the submission process correlates with that found on the title page of the ETD. The library technicians also authorize the name headings and assign LCSH. This entire step takes approximately 30–60 minutes, depending on the difficulty of assigning subject headings.

A library technician uses a modified version of MarcEdit, a homegrown cataloging utility, to map the DSpace Dublin Core metadata to MARC, and export the MARC metadata to the library's online catalog and to WorldCat.[9] MarcEdit saves staff the time of creating MARC records and loading them to OCLC and the local catalog, a time savings of approximately 17 minutes per thesis. Before the final upload to the online catalog and WorldCat, the library technician changes the timestamp to reflect the thesis presentation date rather than the harvesting date and places the student-created keywords into a 653 field. Other edits include editing subfield codes, adding paging and controlling subject headings. The mapping, editing, and export process takes approximately 10 minutes per item and is done at the end of each week.

Because graduate and PhD students are still required to submit one print copy of their research, library staff still must spend time handling and processing T/Ds. The time spent examining and collating T/Ds, however, has been cut in half. Unlike the print T/D workflow, which required T/Ds to change hands five times between five different library employees, the ETD workflow sees items change hands only once, when the binding technician delivers the storage copies to a circulation technician for shelving in the offsite location. This new workflow cut the total processing time by approximately 20 minutes per T/D, resulting in a cost savings of more than $4.00 per T/D.

FUTURE CONSIDERATIONS

Although the current ETD program saves the OSU Libraries both time and money, the workflow continues to be refined in order to further economize the process. Discontinuing full subject analysis and no longer requiring a print copy are two possible changes that could significantly affect workflow.

Discontinuing Full Subject Analysis

As noted in Table 1, more than 50% of the total time spent processing ETDs at OSU is spent assigning LCSH, a practice that appears to be losing favor with academic libraries.[10] Possible reasons for this downward trend include:

- Cataloging of T/Ds is time-consuming and costly.[11]
- T/D subject matter is often narrow and specialized, making it difficult for library staff without the necessary expertise and/or background to assign subject headings accurately.[12]
- T/Ds often contain cutting-edge research and LCSH has not always caught up with newfound knowledge.[13]
- Keyword and full-text searching are alternatives to LCSH.[14]

Reviewing the library literature reveals little agreement as to whether the benefits of assigning LCSH outweigh the expense incurred by libraries continuing to do full subject analysis. Many librarians contend that the improved access LCSH provides is reason enough to continue with the long-standing tradition. Sapon and Hansbrough, for example, discovered that "circulations [are] 58% higher for records with subject headings when compared to those without subject headings."[15]

Circulation statistics for the 445 2004 T/Ds (with LCSH) indicate that T/Ds circulate, on average, a total of one and a half times. Using Sapon and Hansbrough's figures, had these particular T/Ds not had subject headings they would have circulated, approximately, only once. Does the relatively small increase in circulation numbers warrant the time and expense of full subject analysis? No longer adding LCSH would save the Libraries an additional 45 minutes and $14.03 per T/D. A more thorough analysis is necessary before making a decision to discontinue the long-standing library policy.

Electronic Only

Early in the discussions between the Libraries and the Graduate School regarding ETDs, the Graduate School expressed interest in moving to an electronic-only process. Graduate School staff sympathized with students burdened with the high costs associated with printing T/Ds on archival paper, and in some instances, a requirement to provide their academic unit with a bound copy in addition to supplying the Libraries copies. The Graduate School was also eager to economize their own workflow. Not having to receive, handle, and store paper T/Ds would free up both time and space for the department.

Clearly, the Libraries could save time and money by discontinuing the practice of requiring print copies of T/Ds. In 2003, the library decided to

TABLE 2 ETD Electronic Only Comparison

Workflow tasks	ETD (minutes/100 items)	Electronic only (minutes/100 items)
Retrieve theses from Graduate School	60	0
Examine and collate theses	1000	0
Prepare/send dissertations to UMI	30	30
Prepare theses for storage	60	0
Review student-created metadata; authorize names; add descriptive notes	500	500
Assign LCSH	3250	3250
Map/review/export DSpace metadata to OPAC and WorldCat	950	950
Shelve storage copies of theses/dissertations	60	0
Total time spent (minutes/100 items)	6020	4730
Total time spent (minutes/item)	~60	~50
Total processing cost (per item)	$18.87	$15.49
Total annual processing cost (500 items)	$9,435	$7,745

cease sending the archival copies of print T/Ds to a commercial binder and instead began housing these copies in archival quality, acid-free envelopes. This archivally sound practice saves the Libraries roughly $5,000 on binding costs and $4,000 in wages per year. No longer requiring print T/Ds would mean even more savings (see Table 2). The Libraries hesitated to do away with print copies of T/Ds completely. Not only do T/Ds represent the significant research and scholarship of the university, but they are also an important historical record of the OSU research and teaching community. With print T/Ds dating back to 1902, and even much older print materials still in excellent condition, the Libraries felt confident in print T/D archiving practices. Digital preservation was still in its early days, and while the Libraries had developed plans for ensuring long-term access to digital materials, those plans were not yet mature enough to represent a comprehensive strategy for digital stewardship. Print materials need very little attention to endure. Digital objects require constant and continuous attention to remain viable. At the time, the libraries were simply not willing to risk losing these valuable resources until issues surrounding digital preservation management were clearer.

In the past few years, the digital preservation landscape has matured, with many well-trusted sources now providing a variety of resources for institutions and organizations interested in preserving digital objects. The National Archives co-host "Partnerships in Innovation," a conference dedicated to "address the challenges of preserving electronic records";[16] the Library of Congress is partnering with dozens of organizations to investigate and develop models, practices, and tools;[17] and Cornell has developed both an award-winning online tutorial and a series of workshops taught by a

"Who's Who" of digital preservation experts.[18] Because of these efforts, and the Libraries' own increased comfort level with its digital preservation program, the Libraries will again consider no longer requiring a print copy in the coming year.

CONCLUSION

A Council of Library and Information Resources report in 2004 suggested the move from print to electronic journals would result in time savings for Technical Services staff and would allow them to take on new responsibilities.[19] OSU found that the move from print to electronic resulted in time savings in several areas of Technical Services and resulted in Technical Services staff taking on many new and related roles in digital library development. Even though the Libraries continue to require both a print and electronic copy of all T/Ds produced at OSU, since moving to electronic and print, the entire process is completed more efficiently, with fewer steps and fewer staff involved. This is largely the result of using student-created descriptive metadata during the ETD submission process and the automatic generation of MARC records for the online public access catalog and WorldCat. Staff formerly responsible for aspects of the processing and cataloging of T/Ds have been reassigned to prioritized project work including serials retrospective conversion, metadata creation, and gifts processing.

NOTES

1. Networked Digital Library of Theses and Dissertations, "NDLT/D Membership," Networked Digital Library of Theses and Dissertations, http://www.ndlT/D.org/members/index.en.html (accessed May 14, 2008).

2. Edward A. Fox, John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer, "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources," *D-Lib Magazine* 2 (September 1996), doi:10.1045/september96-fox, http://dx.doi.org/10.1045/september96-fox, http://www.dlib.org/dlib/september96/theses/09fox.html (accessed May 14, 2008).

3. Networked Digital Library of Theses and Dissertation, "NDLT/D Membership."

4. Oregon State University Archives and Records Management Program, "General Records Retention Schedule. Theses and Dissertations Records," Oregon State University, http://osulibrary.oregonstate.edu/archives/schedule/student.html#49 (accessed May 23, 2008).

5. Janet Webster et al., "Implementation of a New Library/University Service: Oregon State University's Institutional Repository" (final report, Institutional Repository Task Force, Oregon State University Libraries, Oregon State University, 2005), http://hdl.handle.net/1957/24 (accessed May 14, 2008).

6. Michael Boock, "Improving DSpace@OSU with a Usability Study of the ETD Submission Process," *Ariadne* 45 (October 2005), http://www.ariadne.ac.uk/issue45/boock/ (accessed June 6, 2008).

7. Edward Fox et al., "ETD-MS: An Interoperability Metadata Standard for Electronic Theses and Dissertations," http://www.ndlT/D.org/standards/metadata/current.html (accessed June 6, 2008).

8. Dublin Core Metadata Initiative Libraries Working Group, "DC Library Application Profile (DC-Lib)," http://dublincore.org/documents/library-application-profile/ (accessed June 6, 2008).

9. Terry Reese, Jr. and Kyle Banerjee, *Building Digital Libraries* (New York: Neal-Schuman Publishers, Inc., 2008), 171–3.

10. Robert E. Wolverton, Jr. and Lona Hoover, "Historical Perspectives on the Treatment and Cataloging of Theses and Dissertations," *Technical Services Quarterly* 21 (2004): 9, doi:10.1300/J124v21n03_01, http://dx.doi.org/10.1300/J124v21n03_01.

11. Norma Velez-Vendrell, Jacque Halverson, and Laura Salas-Tull, "Evaluation of a Program for Assigning Subject Headings to Local Theses and Dissertations," *Cataloging & Classification Quarterly* 9, no. 2 (1988): 81–90.

12. Zahiruddin Khurshid, "Improvisations in Cataloging of Theses and Dissertations," *Cataloging & Classification Quarterly* 20, no. 2 (1995): 51–9. doi:10.1300/J104v20n02_04, http://dx.doi.org/10.1300/J104v20n02_04.

13. Brian E. Surratt and Dustin Hill, "ETD2MARC: A Semi-Automated Workflow for Cataloging Electronic Theses and Dissertations," http://handle.tamu.edu/1969.1/588 (accessed July 10, 2008).

14. Cynthia C. Ryans, "Cataloging Theses and Dissertations: An Update," *Cataloging & Classification Quarterly* 14, no. 1 (1991): 83–87, doi:10.1300/J104v14n01_08, http://dx.doi.org/ 10.1300/J104v14n01_08.

15. Richard E. Sapon and Mary Hansbrough, "The Impact of Subject Heading Assignment on Circulation of Dissertations at Virginia Tech," *Library Resources and Technical Services* 42, no. 4 (October 1998): 282–291.

16. The National Archives, "Partnerships in Innovation," The National Archives, http://www.archives.gov/era/presentations/innovations/ (accessed July 11, 2008).

17. Library of Congress, "Digital Preservation," Library of Congress, http://www.digitalpreservation.gov/ (accessed July 8, 2008).

18. Cornell University Library, "Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Problems," Inter-University Consortium for Political and Social Research, http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html (accessed July 8, 2008).

19. Roger C. Schonfeld, Donald W. King, Ann Okerson, and Eileen Gifford Fenton, "The Nonsubscription Side of Periodicals: Changes in Library Operations and Costs between Print and Electronic Formats," Council of Library and Information Resources, http://www.clir.org/pubs/reports/pub130-original%20templates/contents.html (accessed July 12, 2008).

APPENDIX Theses (Graduate) Data Dictionary

Label	Metadata element	Repeatable	Required	Description	Controlled vocabulary
Author	dc.creator	Y	Y	Enter the name of student in the form of: Last name, First name Middle initial, e.g., Doe, Jane K.	N
Title	dc.title	Y	Y	Copy and paste title from the PDF. The first word of the title and any personal, place and corporate names should be capitalized within the title. Other words in the title should not be capitalized.	N
Advisor	dc.contributor.advisor	Y		Enter the name of thesis adviser in the form of: Last name, First name Middle initial (when available), e.g., Doe, Jane K.	N
Committee Member	dc.contributor.committee member	Y	N	Enter the name of Committee Member(s) in the form of: Last name, First name Middle initial (when available), e.g., Doe, Jane K.	N
Date	dc.date.issued	N	Y	This should be the date that appears on the title page or equivalent of the work. Use YYYY, e.g., 2005.	N
Abstract	dc.description.abstract	Y	N	Copy and paste the abstract from the PDF.	N
Description	dc.description	Y	N	Library enters the following Description fields: 1. Graduation Date: YEAR	N
Keywords	dc.subject	Y	N	Enter keywords on the topic of the thesis.	N
Keywords	dc.subject.lcsh	Y	N	Library will add Library of Congress Subject Headings.	N
Degree Name	thesis.degree.name	Y	Y	Enter degree name. Library enters complete degree name: e.g., Master of Science (MS) *in Electrical and Computer Engineering*	Y
Degree Level	thesis.degree.level	Y	Y	Enter level of education associated with the document.	Y
College	thesis.degree.discipline	Y	Y	Enter the name of the college or department.	Y
Granting Institution	thesis.degree.grantor	Y	Y	Enter institution granting the degree associated with the work.	Y
URI	dc.identifier.uri	N	Y	Generated automatically. Contains a permanent handle URL.	
	dc.type	N	Y	Select: Thesis	Y
	dc.language.iso	N	Y	Select: English (United States)	Y

(*Countinued on next page*)

APPENDIX Theses (Graduate) Data Dictionary (*Continued*)

Label	Metadata element	Repeatable	Required	Description	Controlled vocabulary
Fields generated automatically by DSpace:					
	dc.date.accessioned	N	Y	Generated automatically.	
	dc.date.available	N	Y	Generated automatically.	N
	dc.description.provenance	N	Y	Generated automatically. Contains information about document submitter, date submitted, number of bitstreams, file name, byte size, and checksum.	N
	dc.description.provenance	N	Y	Generated automatically. Contains information about document approver, date approved, number of bitstreams, file name, byte size, and checksum.	N
	dc.description.provenance	N	Y	Generated automatically. Contains information about the date the document was made available in archive, number of bitstreams, file name, byte size, and checksum.	N
	dc.format.extent	N	Y	Generated automatically. Contains number of bytes.	N
	dc.format.mimetype	N	Y	Generated automatically. Contains mimetype of file.	N
Appears in Collections		N	Y	Generated automatically.	N