6-26-2009

Repurposing ProQuest Metadata for Batch Ingesting ETDs into an Institutional Repository

Shawn Averkamp
University of Iowa

Joanna Lee
University of Iowa

Code4Lib journal, 7 (2009). http://journal.code4lib.org/articles/1647

code{4}lib JOURNAL

Issue 7, 2009-06-26 ISSN 1940-5758

Repurposing ProQuest Metadata for Batch Ingesting ETDs into an Institutional Repository

This article describes the workflow used by the University of Iowa Libraries to populate their institutional repository and their catalog with the data collected by ProQuest UMI Dissertation Publishing during the submission of students' theses and dissertations. Re-purposing the metadata from ProQuest allowed the University of Iowa Libraries to streamline the process for ingesting theses and dissertations into their institutional repository The article includes a discussion of the benefits and limitations of the workflow described.

by Shawn Averkamp and Joanna Lee

Introduction

The University of Iowa Libraries has recently established an institutional repository (IR) for archiving a broad range of scholarly output including graduate student theses and dissertations. We expect the quantity of theses and dissertations submitted electronically to increase as The Graduate College begins encouraging electronic submission over traditional print submission. Therefore, we needed to create an efficient workflow for batch ingesting this content into our IR (Iowa Research Online) [1]. The University of Iowa currently uses the ProQuest UMI Dissertation Publishing service to handle processing of both print and electronic theses and dissertations. To submit electronic theses and dissertations (ETDs) to ProQuest, students complete a web form and upload their documents. This data is later returned to the Libraries via ftp as XML metadata, a PDF of the thesis or dissertation, and any supplementary files. We developed an XSLT stylesheet to convert the ProQuest XML metadata to an upload-ready XML schema. While it is possible to harvest this metadata in XML from catalog records in WorldCat, we chose to use the ProQuest metadata for a variety of reasons, foremost being that we can make the electronic access copies available before local MARC catalog records are created, and then generate brief MARC records for our local catalog.

In this article, we will present a detailed description of this process including its benefits and limitations. Our repository is hosted on Digital Commons a platform developed by bepress (The Berkeley Electronic Press), but the workflow we will outline, summarized by Figure 1, could

easily be adapted by institutions using other repository platforms such as DSpace [2]. As we developed this approach, we tried to integrate and streamline existing workflows and repurpose metadata as much as possible to avoid manual processes. For example, the ProQuest data was an attractive source because the metadata was robust and the files were readily available through ProQuest's FTP delivery. The MARC records we generate, while still requiring some manual review, could enhance the previous workflow for handling ETDs in which individual records were created manually and added to the local catalog. We include the annotated XSL files we developed for others to use and adapt[3].

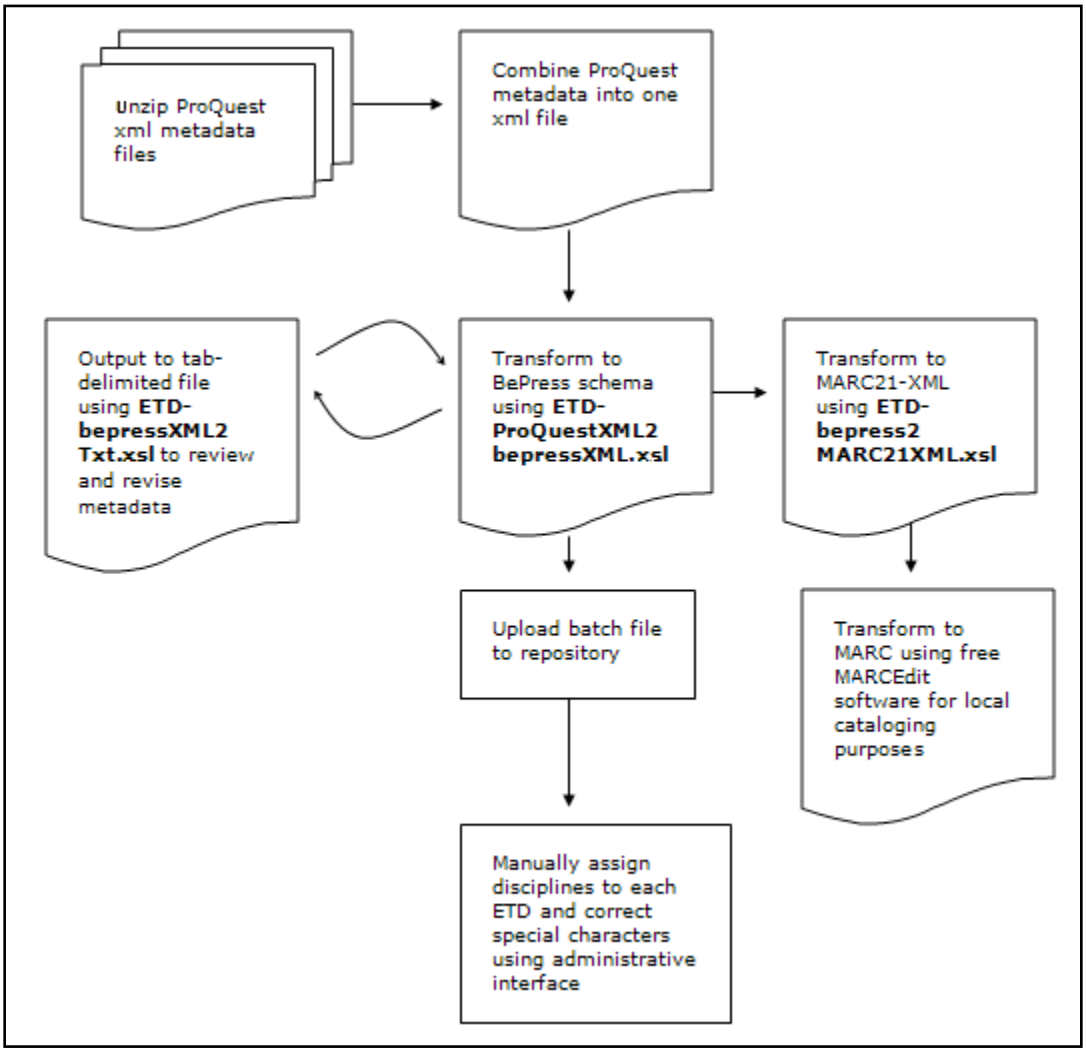

Figure 1: ETD workflow

Process

Unzip ProQuest XML metadata files

ProQuest submits compressed folders of ETD XML metadata and PDF files to The University of Iowa via FTP. The first step is to unzip all of the ETDs that need to be uploaded to Iowa Research Online and make sure the PDF files are stored in a publicly accessible directory so that bepress can automatically pull them in during the batch upload.

Combine ProQuest metadata into one XML file

ProQuest returns metadata about each ETD in a separate XML file. Because we needed to build one batch file describing all of the ETDs, the next step is to combine all of the XML files into one XML file.

Instructions for Combining Multiple XML Documents into One XML Document using Microsoft Windows.

1. Copy filepaths of files to be combined into a Microsoft Excel spreadsheet column (On Vista, hold shift while right-clicking, select “copy path.”)
2. In Excel, remove the topmost common directory from filepaths. (Use find and replace to remove, for example, “`L:\etd.lib.uiowa.edu\`“)
3. Add a “+” to each filepath. (In the next column, B1, enter concatenate formula: =*`concatenate(A1,"+")`* then copy formula down the column for the rest.)
4. Copy column B into a Word document (using Paste Special–unformatted text).
5. Remove line breaks in the Word document (using Find and Replace to remove “`^p`“).
6. Delete last “+” from this text string.
7. In command line, change to the desired drive. (For example, if you are working with files on the L drive, enter: “`L:`“)
8. Change to topmost common directory (the one you removed from the file paths in step 2). (For example, enter: “`chdir etd.lib.uiowa.edu`“)
9. Use the “copy” command to combine the files (enter: “`copy`“) then paste in the string you created in word, then a space, then name the destination file. (For example, “`ETDmetadata.xml`“)
10. Open your destination file in Notepad and using Find and Replace, remove all of the “`<?xml…>`” headers.
11. Paste the header back in at the top of the document. Just below, enter a top-level “`<xml>`” element and close at the end with “`</xml>`“.

Transform to bepress schema using ETD-ProQuestXML2bepressXML.xsl

Because of bepress' minimal metadata schema, crosswalking between the two schemas was fairly straightforward. Elements that are not represented explicitly in bepress' schema can be represented with a custom field element. While this solution offers flexibility in mapping from ProQuest to bepress, mapping from the custom fields in bepress to Dublin Core is currently somewhat limited; bepress is working on their metadata export tools and will be offering more options in the future. (See Table 1 for complete crosswalk from ProQuest to bepress schema.)

ProQuest (source)	bepress (output)	Format/Value	Notes
xml	documents		root
DISS_submission	document		item
DISS_description/DISS_title	title		Outputs first letter of each word in caps
DISS_authorship/DISS_author/DISS_name/DISS_surname DISS_authorship/DISS_author/DISS_name/DISS_fname DISS_authorship/DISS_author/DISS_name/DISS_middle	authors/author/lname authors/author/fname authors/author/mname		
	authors/author/institution	The University of Iowa	
DISS_content/DISS_abstract/p	abstract/p		
DISS_description/DISS_dates/DISS_comp_date	publication-date	ISO 8601 (yyyy-mm-dd)	Completion date from ProQuest record. Only year is set to display in our repository, but since bepress requires the full date, we've defaulted to Jan 1.
DISS_description/DISS_categorization/DISS_keyword	keywords/keyword		
DISS_content/DISS_binary	fulltext-url		Filename concatenated with local location on server staging area
DISS_description/DISS_degree	fields/field/@name="degree_name"/value		
DISS_description/DISS_institution/DISS_inst_contact	fields/field/@name="department"/value		
DISS_description/DISS_advisor[1]/DISS_name/DISS_fname	fields/field/@name="advisor[1]"/value		First three advisors captured

DISS_description/DISS_advisor[1]/DISS_name/DISS_surname DISS_description/DISS_advisor[1]/DISS_name/DISS_middle			
	label	Item node position + total of ETDs already uploaded	Forces bepress article id on each ETD in order to construct the access URL before uploading

Table 1: ProQuest Schema to bepress Schema Metadata Mapping

Now we can transform the XML file from ProQuest's schema to bepress's general schema for importing material using the transformation we developed, ETD-ProQuestXML2bepressXML.xsl [4]. There are a few aspects of our XSL file worth noting. We use a function to build the filepath of each PDF for the fulltext-URL field based on the filename of each ETD and the root directory where the file is saved. During the batch uploading process, bepress will use this path to pull in each file. Our code also normalizes the degree_name field to keep those values consistent (for example, M.A. will change to MA; phd will change to PhD) and uses a function to change the all-caps title field values and the variable name field values of the ProQuest metadata to title case (only the first letter of each word is capitalized).

In order to control and predict the future URL of each ETD when it is uploaded, we forced incremental integers onto each record in bepress's label field. For example, an ETD with a label value of "75″ will be uploaded to http://ir.uiowa.edu/etd/75. bepress automatically generates a label for each document, but we chose to force the label so that we can easily generate the URL of each record for local cataloging purposes in a later transformation. Each time we transform a new batch of ETDs to bepress's schema, we must change the base integer in the label field of ETD-ProQuestXML2bepressXML.xsl to ensure that each ETD receives a unique label and thus, a unique URL. For example, if there are 220 ETDs loaded into Iowa Research Online, we must start the next batch at integer 221.

It is possible to create a unique label for each ETD without relying on a single base integer that must be reset before each transformation; the bepress schema supports any string in this field. For example, combining an integer and another field, such as the author's last name, will likely generate a unique value for the label field and result in a URL such as http://ir.uiowa.edu/etd/chang13. However, on bepress's recommendation, we chose to use an integer value alone to make the ETDs easier to manage in the administrative interface; sorting by label reflects the order in which documents were uploaded. It also keeps the format of the resulting URL consistent with other documents in our repository (http://ir.uiowa.edu/series-name/integer).

Output to tab-delimited file using ETD-bepressXML2Txt.xsl to review and revise metadata

To make it easier to review the transformed metadata before batch-loading, we developed another transformation to reformat it as a tab-delimited file. Opened in a spreadsheet, it is easy to check the metadata for errors and make associated changes in the transformed XML.

Upload batch file to repository

Now we simply upload our batch file to Iowa Research Online.

Transform to MARC21-XML using ETD-bepress2MARC21XML.xsl

Now the ETDs are available in the repository and discoverable in the local catalog through a pipe from Iowa Research Online, but they do not have MARC records in the local catalog. At this point, The University of Iowa plans to continue building MARC records for the ETDs, in part to easily maintain them in OCLC WorldCat. To make local cataloging easy, we developed another transformation to reformat the metadata from our final bepress batch XML file to MARC21XML. Because we forced a label on each ETD during the transformation to bepress's schema, we are able to predict and build the final URL of each ETD in our MARC21XML.

Create MARC records using MarcEdit

The final step in our workflow is to use an existing transformation available from MarcEdit, a free program developed by Terry Reese at Oregon State University, to generate MARC records from our MARC21-XML [5].

Benefits

The main benefit of using the ProQuest metadata is that we are able to provide public access to ETDs sooner than we would have been able to by harvesting OCLC metadata post-cataloging. IR metadata records created from a ProQuest schema to bepress schema transformation, while imperfect, can serve as access records while the ETDs await local cataloging for the library catalog. Although MARC records do not yet exist in the local catalog, a pipe from our bepress IR to our library's federated search system (Ex Libris' Primo) allows ETDs to be discoverable in the local collection environment. When ETDs eventually receive local cataloging treatment, minor errors in title case and special characters can be corrected in the IR metadata [6]. (While the creation of duplicate records in both the IR and the local catalog may seem redundant and potentially confusing to the federated search user, current local cataloging guidelines do not allow for the substitution of an IR metadata record for a full MARC catalog record.)

To expedite local cataloging of ETDs, we created a transformation that could be used to automate some of the process of brief record creation. (This step has not yet been approved for our local cataloging workflow, but we include it for the benefit of those wishing to streamline their own workflows.) ProQuest metadata that has been transformed to the bepress schema can then be transformed into MARC21XML, which can in turn be transformed into MARC-21 brief records using the MarcEdit tool (developed by Terry Reese at Oregon State University, see Resources, below). Brief records were previously created manually. Using local thesis and dissertation cataloging guidelines and the Networked Digital Library of Theses and Dissertations' (NDLTD) ETD-MS interoperability metadata standards, we created a transformation to convert our bepress upload XML (after manual edits have been completed) to MARC21XML [7]. This transformation captures most datafields, but a few must be entered manually, specifically physical description (300ab) and topical subject headings (650ax). Also, the University of Iowa Libraries adds several local fields that must be populated manually. As noted earlier, the title statement (245:10abc) case must be normalized manually and the abstract (520) must be checked for mistranslated diacritics.

Using MarcEdit, a free application for editing and transforming MARC records, we convert the MARC21XML to a MARC file (.mrc). This file can then be imported into the integrated library system (ILS) where certain fields are populated automatically, and the records can be cleaned up and fleshed out.

bepress	MARC	Format/value	Notes
	LDR	^^^^^ntm^ 22^^^^ ua 4500	
	005		auto-generated in Aleph
	006	m^^^^^^^^d^^^^^^^^	Additional fixed data: electronic resource
	007	cr^n|||||||||	Physical description fixed field: electronic resource
	008/15-17	xx	Publication place: no place of publication – i.e. unpublished
	008/23	s	Form of item: electronic
	008/24-27	m	Nature of contents: theses
	040ac	NUI	Cataloging source
authors/author/lname authors/author/fname authors/author/mname	100:1a	[lname], [fname] [mname]	Main Entry
title	245:10ab		Title is split into title and subtitle at ':', '?', or '?:'

authors/author/lname authors/author/fname authors/author/mname	245:10c	by [fname] [mname] [lname].	Usually consistent with title page
	245:10h	[electronic resource]	General material designation: electronic resource
publication-date	260c	yyyy.	
fields/field/@name=”advisor[1]”/value fields/field/@name=”advisor[2]”/value fields/field/@name=”advisor[3]”/value	500a	Thesis supervisor: [advisor[1]] Thesis supervisor: [advisor[2]] Thesis supervisor: [advisor[3]]	Thesis advisor(s)
fields/field[@name='degree_name']/value	502a	Thesis ([degree name]) – University of Iowa, 2008.	Dissertation Note
abstract	520:3a 520:8a		Paragraph breaks are removed. Local guidelines allow 2000 characters for 520:3a field. Remaining characters are entered into 520:8a fields.
	538a	Mode of access: World Wide Web.	System details notes (As recommended by NDLTD.)
	538a	System requirements: Adobe Reader.	
fields/field[@name='advisor1']/value fields/field[@name='advisor2']/value fields/field[@name='advisor3']/value	720a	[advisor[1]] [advisor[2]] [advisor[3]]	Added entry – uncontrolled name. (As recommended by NDLTD. Univ. of Iowa Libraries uses 700:10, local added entry field, instead.)
	720e	advisor.	
label	856:40u	http://ir.uiowa.edu/etd/[label]	Concatenation of IR collection directory path and bepress XML label

Table 2: bepress Schema to MARC21XML Metadata Mapping

Limitations

There are some limitations to our workflow that require manual corrections before and after the batch file is uploaded. The first involves the title field. When students submit a thesis or dissertation, the Thesis Manual of the Graduate College dictates that the title must be formatted

in all caps. The ProQuest metadata inherits this convention, so we generally have all-caps title fields to work with. We chose to reformat them as title-case (only the first letter of each word is capitalized). From a cataloging perspective, the resulting output is not ideal, but since proper nouns are not differentiated in the all-caps of the original metadata, using title-case was a good, scalable compromise. A side-effect of this decision is that any acronyms in the titles will be incorrect (for example, DNA is changed to Dna). As a final step before uploading, we make manual corrections to the titles in the XML batch file. This manual process can be eased by reviewing a tab-delimited text transformation of the metadata in MS Excel (ETD-bepressXML2Txt.xsl). In a future revision of our style sheet, we would like to address this problem by creating a list of common acronyms and checking the titles against them. Any strings in the title that match an acronym on the list would remain capitalized.

The source data for publication-date field is a bit problematic, too. We chose to map the publication-date field from the DISS_comp_date field. However, this Proquest field generally only contains a year (formatted as yyyy). bepress requires the publication date to include month, day, and year and conform to ISO 8601 (yyyy-mm-dd), so our transformation adds "01″ for month and date (2008 becomes 2008-01-01).

Other limitations are due to bepress' schema. bepress' batch loading feature only supports the Latin-1 character set, so any characters outside of that set need to be corrected after the ETDs are uploaded. Though you can add many custom fields, currently bepress' schema cannot be extended to include the discipline in which an ETD should be categorized. Without this discipline information, the ETDs will not be visible when users browse by discipline, either on the homepage of the repository or in bepress' cross-repository ResearchNow! portal. Therefore, each ETD must be manually categorized within the administrative interface of the repository. bepress is planning to support discipline mapping in a future revision of their batch-loading schema. In addition, for bepress fields that we have limited by a controlled drop-down list (such as department), any values with special characters such as ampersands (for example, Electrical & Computer Engineering) will not match correctly during the batch-loading process. To prevent this, we asked bepress to replace all ampersands in controlled fields to "and" and we use a function in our transform (ETD-ProQuestXML2bepressXML.xsl) to change the "&" symbols in any controlled value fields to "and."

Tools

- Markup language: XML 2.0
- XML editing and transformations: <oXygen/> XML Editor 10.0
- Processor (via <oXygen/>): Saxon-B 9.1.0.3

- Combining multiple XML documents: Notepad; Microsoft Office Word 2007; Microsoft Office Excel 2007
- Viewing tab-delimited files: Microsoft Office Excel 2007
- Generating MARC files (.mrc) from MARC21XML: MarcEdit [5]

Other resources

There are several other resources that might be helpful for institutions interested in repurposing ETD metadata for their repositories. Michael Witt and Mark Newton at Purdue University have produced an outstanding tutorial about transforming EndNote metadata, "Preparing Batch Deposits for Digital Commons Repositories [8]."

On IUScholarWorks, Randall Floyd from Indiana University Libraries describes a workflow for ingesting ProQuest/UMI metadata and ETDs into a repository built on DSpace [9].

Conclusion

In the future, ProQuest may change the way they structure or deliver their metadata. Other factors could also change, such as how students are required to submit ETDs and how bepress wants data structured for import. While at some point we may need to repurpose metadata from other sources or revise our transformations, we have developed a successful workflow for efficiently ingesting ETDs. In addition, developing our transformations was a great introduction to XSL; we are now applying these skills to target other digital resources for our repository and to repurpose metadata for other digital library applications.

Links

1. Iowa Research Online (IRO) — http://ir.uiowa.edu
2. bepress — http://www.bepress.com/
3. Download the University of Iowa Libraries' XSLT files
4. bepress document-import schema — http://www.bepress.com/document-import.xsd
5. MarcEdit — http://oregonstate.edu/~reeset/marcedit/html/index.php
6. Ex Libris Primo — http://www.exlibrisgroup.com/category/PrimoOverview
7. NDLTD — http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html

8. Witt, M. & Newton, M. (2008). Preparing batch deposits for Digital Commons repositories. *Purdue E-Pubs. Library Research Publications*. Paper 96. http://docs.lib.purdue.edu/lib_research/96/

9. Automated Electronic Thesis and Dissertations Ingest. *IUScholarworks*. http://wiki.dlib.indiana.edu/confluence/x/01Y

Acknowledgments

We would like to thank Wendy Robertson, Digital Resources Systems Librarian, University of Iowa Libraries, for her contributions to the ETD project, her guidance on cataloging practices and local workflows, and her editorial suggestions.

About the Authors

Shawn Averkamp (shawn-averkamp@uiowa.edu) and Joanna Lee (joanna-lee@uiowa.edu) are Digital Projects Librarians at The University of Iowa Libraries Digital Library Services.