Notes on Operations

Leveraging Author-Supplied Metadata, OAI-PMH, and XSLT to Catalog ETDs

A Case Study at a Large Research Library

Ken Robinson, Jeff Edmunds, and Stephen C. Mattes

Most academic theses and dissertations are now born-digital assets (i.e., electronic theses and dissertations). As such, they often coexist with author-supplied metadata that has the potential for being repurposed and enhanced to facilitate discovery and access in an online environment. The authors describe the evolution of the electronic thesis and dissertation (ETD) cataloging workflow at a large research library, from the era of print to the present day, with emphasis on the challenges and opportunities of harvesting author-supplied metadata for cataloging ETDs. The authors provide detailed explanations of the harvesting process, creating code for the metadata transformations, loading records, and quality assurance procedures.

In August 2013, the Cataloging and Metadata Services Department of the Pennsylvania State University Libraries created the Digital Access Team in response to the need to devote more resources to the management of metadata for digital assets.[1] One of the team's primary activities is repurposing metadata from existing MARC records in Penn State's online catalog, The CAT, for digital collections in CONTENTdm and other platforms.[2] The team also works closely with the Library Technologies Department to repurpose MARC records in The CAT for mass digitization partnerships, such as HathiTrust and the Internet Archive.

The team began looking at repurposing metadata from other platforms for use in The CAT in October 2013. One promising source of metadata was Penn State's electronic theses and dissertations (ETDs) server.[3] Metadata for each ETD is available in unqualified Dublin Core (DC) format and can be harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[4] An important tool for harvesting this data is MarcEdit, a freely available metadata editing suite designed by Terry Reese.[5] MarcEdit provides many default crosswalks for mapping between multiple metadata schemes. These schemes can be customized for local metadata harvesting. MarcEdit also includes a tool for harvesting metadata from sites that have implemented or use OAI-PMH.

This paper describes the Digital Access Team's efforts to design an ETD cataloging workflow by harvesting author-supplied metadata using a customized DC-to-MARCXML Extensible Stylesheet Language Transformation (XSLT) crosswalk in MarcEdit to create a file of Resource Description and Access (RDA) MARC records for batch loading into The CAT.[6] The history of thesis cataloging at Penn State is described, including the transition to cataloging ETDs, and how the new harvesting method has improved access to ETDs while simultaneously freeing up staff time. Examples of MARC records for ETDs before and after the new procedure was implemented are provided, and time savings are quantified

Ken Robinson (kjr106@psu.edu) is a Digital Access and Metadata Specialist in Cataloging and Metadata Services at Penn State University Libraries. **Jeff Edmunds** (jhe2@psu.edu) is Digital Access Coordinator in Cataloging and Metadata Services at Penn State University Libraries. **Stephen C. Mattes** (smattes@uoregon.edu) is an Information Specialist in the Oregon Career Information System at the University of Oregon.

Manuscript submitted May 26, 2015; returned to author September 29, 2015 for revision; revised manuscript submitted November 23, 2015; returned to author for minor revision February 4, 2016; accepted for publication March 25, 2016.

on the basis of studies conducted over a twelve-month period (three semesters). The paper also describes in detail the mappings created to harvest the metadata, the customizations made to the XSLT crosswalk, and the steps taken to ensure that the metadata batchloaded into The CAT is of sufficiently high quality.

Literature Review

Literature addressing the harvesting of ETD author-supplied metadata for creating MARC records for online catalogs is somewhat sparse, although efforts date back as far as 1999. Early harvesting strategies used Perl scripts. Sharretts, Shieh, and French described how the University of Virginia Library's pilot project using the Unix command-line utility Grep to extract bibliographic data from thesis PDF title pages and how it evolved into a series of Perl scripts that ran when a student submitted an ETD online.[7] Surratt and Hill described a similar process at Texas A&M University using a Perl script called ETD2MARC that took advantage of the open source MARC::Record Perl module.[8]

As OAI-PMH became more common, libraries began using this protocol to harvest ETD author-supplied metadata. Reeves described a process that Library and Archives Canada (LAC) used to harvest metadata with OAI-PMH queries that retrieved ETD Metadata Standard (ETD-MS) records for ETDs submitted from various Canadian universities in the Thesis Canada Portal.[9] Using this method, LAC had a cost savings of $95,000 in the 2006–7 fiscal year and expected progressively larger savings as more Canadian universities implemented ETD submission programs. McCutcheon et al. described an elaborate process at Kent State University in which a Perl script called ETDcat ran when it received an automatically generated notification from the OhioLINK ETD Center that an ETD had been submitted.[10] The script constructed an OAI-PMH query and retrieved the metadata as an ETD-MS record, which was converted into a MARC record using the MARC-Perl library.[11] Reese documented efforts made by Oregon State University (OSU) to harvest ETD metadata using MarcEdit's Metadata Harvester, which sent an OAI-PMH query that retrieved unqualified DC records.[12] A specialized XSLT crosswalk derived from a default DC to MARC crosswalk that is part of the MarcEdit installation was used to convert the records into MARCXML. Boock and Kunda also described the OSU experience, but focused more on workflow changes and cost savings.[13] They noted a time savings of seventeen minutes for cataloging each thesis using the new method described by Reese. Deng and Reese described further attempts of XSLT crosswalk customization at both OSU and Wichita State University for OAI-PMH ETD metadata harvests.[14] Bower, Courtois, and Turvey-Welch presented a similar OAI-PMH harvesting process for ETDs at Kansas State University.[15] Walsh provided an overview of metadata repurposing using XSLT and gave a user case showing a step-by-step process for harvesting author-generated metadata for ETDs using MarcEdit.[16]

Another avenue for acquiring ETD author-supplied metadata was to repurpose data supplied by ProQuest. Averkamp and Lee documented how the University of Iowa Libraries transformed ProQuest XML files using XSLT to create metadata that could be loaded into their online repository and was used to create MARC records for their online catalog.[17] Middleton, Dean, and Gilbertson described how the University of Arkansas Libraries used ETD author-submitted metadata supplied by ProQuest in MARC format.[18]

Although the literature addressed multiple ways to acquire ETD author-supplied metadata, the variable and often substandard quality of this metadata arose as a common theme. McCutcheon gave a good summary of the issues and noted that "the descriptive record created by automatic harvesting is only as good as the quality of the author-supplied metadata, which varies from author to author."[19] Metadata quality issues included representation of scientific symbols and diacritics, separation of titles from subtitles, nonfiling characters in the title proper, capitalization, management of whitespace, spelling, and other data entry errors.

History of Thesis and Dissertation Cataloging at Penn State

Cataloging of print theses and dissertations (TDs) at Penn State has historically been minimal level and formulaic. Catalog records generally consisted of the full title, author, date of issuance, a pagination count, degree type, and graduate degree program (in a local MARC 699 field). Library of Congress Subject Headings (LCSH) were assigned until 1964, though the headings were generally broad in scope. From 1965 until 1974, LCSH were added only when a personal name, corporate name, or title of a work were present in the TD title. Beginning in 1975, full subject analysis was performed and LCSH was assigned only for TDs containing the term *Pennsylvania* or a local Pennsylvania name (such as a town or county) in the title. This practice has continued to the present. With this workflow, the average thesis required ten to fifteen minutes to catalog, with an additional five to ten minutes per thesis if referred for subject analysis.

Such a relatively minimalist approach was designed primarily as a balance between providing sufficient access for TDs while minimizing the amount of time spent on complicated subject analysis for what are generally very narrow and specialized subject areas. Special Collections Team catalogers perform full subject analysis for any TDs added to Penn State's Special Collections Library.

Penn State University Libraries initiated a pilot project in collaboration with the Graduate School, Information Technology Services, and Digital Library Technologies in the fall of 1998 to investigate the possibility of allowing theses and dissertations to be submitted and archived electronically. The Graduate School began accepting ETDs in 2000. Penn State originally used ETD-db, an open-source ETD database developed at Virginia Tech.[20] The current ETD application is based on ETD-db and developed with Django, an open source web application framework and MySQL, an open source relational database management system.[21] Records provided in OAI-PMH feeds are currently in unqualified DC format.

Cataloging of ETDs began in 2004. The existing minimalist approach to cataloging TDs was used as a foundation for cataloging ETDs. Electronic aspects were added to the catalog records (MARC 006, 007, 538, and 856), added entries for thesis advisors were included for the first time (in MARC 700), and author-supplied keywords were added to MARC 653. Because ETDs were accessible online, catalogers began copying much of the data from the online record found in Penn State's ETD database and from the title page of the ETD's PDF file. This data was pasted into a MARC record in The CAT. To save time, ETD cataloging was supplemented with a series of Macro Express macros for data repeated in every MARC record.[22] Repeatable data included fixed field data elements in MARC 006, 007, nd 008, MARC 260 (Publication, Distribution, etc.), and MARC 538 (System Details Note). Because thesis titles in PDF files were sometimes entirely capitalized, copying and pasting proved to be as time-consuming as typing the title from scratch. The Digital Access Team's programmer created a script using AutoIt (a freeware automation scripting language) that adjusted the capitalization for pasting into The CAT.[23] Finally, an additional script was created to convert a bulleted list of keywords into a single MARC 653 field.[24] Using this approach, the average ETD required between five and ten minutes for cataloging. Cataloging an ETD generally took about half as long as cataloging a print thesis, a time savings due primarily to the efficiency gained through the copying and pasting of data.

Old Workflow, 2004–14

After receiving a list of ETDs from the Graduate School thesis office each semester, the thesis cataloger cataloged each ETD individually. Starting with a blank template in the local SirsiDynix Symphony ILS, the cataloger used macros line-by-line to fill in constant fields (fixed fields, 006, 007, 040, 260, 538). The cataloger transcribed or copied the title, author, degree type, advisor(s), and thesis department as they appeared on the ETD document. The cataloger took metadata from the ETD server page when it did not appear in the document, such as keywords for the 653 field. The URL provided in the 856 field led to the splash page for the individual thesis.

The cataloger provided local authority control for ETD authors and advisors by searching the local catalog for any previous works by the author or advisor and using the form of name found. If no previous works were found, the name was entered in MARC 100 and 700 using the usage found on the ETD. The cataloger added MARC 246 fields for title variations, such as an alternate form for hyphenated words, or discrepancies between the title on the PDF and that on the ETD server page. The cataloger also added pagination information in the 300 Physical Description field, and checked for additional files to list in 300 subfield $e, such as audio or video files. Typically, the cataloger would spend the bulk of a month (100–160 person-hours) cataloging 300–400 ETDs after each semester.

With a shrinking staff, competing demands for time, and new priorities (such as the creation of metadata for digital projects), Cataloging and Metadata Services felt the time was right to transition from a largely manual, title-by-title process for cataloging ETDs to a more automated, batch approach that leveraged the power of harvesting author-supplied metadata.

Harvesting Metadata versus Records from ProQuest

All doctoral dissertations at Penn State are submitted to ProQuest/UMI Dissertation Publishing for microfilming.[25] This arrangement, initiated for the purpose of preservation, has been in place for over fifty years. Penn State does not submit master's theses to ProQuest, but authors may submit their master's thesis abstracts to ProQuest's Master's Abstract program. Undergraduates in Penn State's Schreyer Honors College are required to complete an undergraduate honors thesis. These are not submitted to ProQuest.

All doctoral dissertations and master's theses are currently submitted to Penn State's ETDs site and all undergraduate honors theses are currently submitted to Penn State's Electronic Honors Theses (EHTs) site.[26] Both sites are OAI-PMH compliant. Because metadata are readily available and can be harvested for all ETDs and EHTs and ProQuest only has metadata for doctoral dissertations, the Digital Access Team chose to harvest this data locally.

Harvesting Metadata: The Process

Metadata are harvested from Penn State's ETD server (etda.libraries.psu.edu) using MarcEdit's Metadata Harvester. The harvest process includes an XSLT crosswalk that transforms the DC data output by the OAI-PMH feed into

RDA-compliant MARC records.

The server name and query date parameters are entered in the Server box, for example,

> https://etda.libraries.psu.edu/oai/?verb=ListRecords&from=2014-04-01&until=2014-09-15&metadataPrefix=oai_dc

The Metadata type is set to "Dublin Core." Other options include OAI MARC, MODS, and MARC21XML.

The Crosswalk is set to the XSLT file locally customized to provide RDA-compliant MARC data in conformance with local standards for ETD metadata:

> C:\Program Files\MarcEdit 6\xslt\ETD_OAIDCtoMARCXML-rev5.xsl

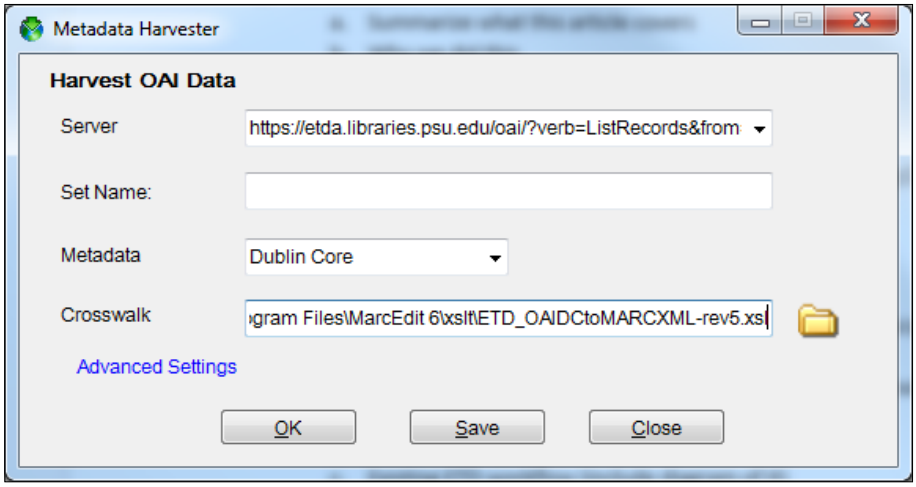

Figure 1. MarcEdit's Metadata Harvester

Clicking on "OK" initiates the harvest, which usually takes only a few seconds for batchloads containing hundreds of records. The file of harvested records then automatically opens in MarcEdit's editor window. At this point, the MARC data can be further manipulated as needed for quality assurance. The edited .mrk file (a MarcEdit file format that is readable and easily editable by a human) is ultimately compiled into a MARC file in .mrc format for loading into the local ILS where the records are again spot-checked for quality to verify that aspects of the records more readily noticeable in the public Webcat interface are in fact correct and display as expected.

Tweaking the Dublin Core Mappings

The metadata available from the ETD server via the OAI-PMH harvest are largely author-supplied (i.e., input by authors at the time they upload their ETDs to the site). These data elements are internally mapped to DC elements. Since DC is a much less precise framework than AACR2 or RDA, the first hurdle faced was mapping the vagaries of DC to the precision of RDA expressed in MARC.

The data elements available on the ETD site include "Graduate Program" and "Keywords." Both were originally mapped to the DC element *subject*, which meant that the out-of-the-box XSLT transformation Penn State used as a test (OAIDCtoMARCXML.xsl) transformed both elements to the MARC 690 field, a local subject access field. Penn State's practice for ETDs has been to distinguish between keywords, which were manually input in MARC 653 (Index Term—Uncontrolled) fields, and graduate program data, which was input using MARC 699, a local subject access field specifically for collating theses by graduate program in The CAT. The solution was to change the mapping of the Graduate Program ETD element to a DC element not used elsewhere in the data, *coverage*, and edit the .xsl file to output the DC element *coverage* as MARC 699. The .xsl file was also edited to output *subject* not as the default 690 but as MARC 653 instead.

A similar conundrum existed for ETD data elements originally not mapped to DC. Neither "Degree" nor "Committee" data was mapped to DC, and therefore not output in the harvest. Degree information needed to be mapped to the MARC 502 (Dissertation Note) and committee member data (i.e., personal names) mapped to MARC 700 fields (Personal Name Added Entries) with relationship designators. The solution was to map degree to a DC element not used elsewhere in the mappings, *relation*, and then edit the .xsl file to map DC element *relation* to MARC 502. Because this mapping is not standard, Penn State Libraries' Digital Content Strategist (who remains in close contact with stakeholders at the Graduate School, the Honors College, and other institutions) was consulted to ensure that mapping the data in this way would not disrupt existing harvest workflows or negatively impact other potential harvesters of Penn State's data via the OAI-PMH feed. The DC element *contributor* had not been used in the mappings, so Committee was mapped to *contributor* and the XSLT file was edited to output this data in 700 fields.

Finally, there was the issue of MARC data completely absent from the ETD author-supplied data and handled imprecisely or not at all by DC: 006, 007, 008 (Fixed Length Data Element fields), 040 (Cataloging Source), etc. Much of this could be added to the records with MarcEdit following harvesting, but customizing the XSLT transformation

allowed us to add this data as part of the harvest itself.

Table 1 shows the ETD data elements available for harvest and their corresponding DC and MARC mappings.

DC data output from the OAI-PMH harvest was transformed and correctly formatted to RDA/MARC using a customized XSLT file.

Table 1. ETD, Dublin Core, and MARC Mappings

ETD Data Element	Dublin Core	MARC
Author	creator	100
Email	N/A	N/A
Graduate Program	coverage (previously mapped to "subject")	699
Degree	relation (previously not mapped)	502
Document Type	type	N/A
Date of Defense	date	264 $c, 008 Date 1
Committee	contributor (originally not mapped)	700
Availability	rights (originally not mapped)	506
Title	title	245
Abstract	description	520
Keywords	subject	653
Files	identifier	856

XSLT Customization

Following initial testing of MarcEdit's *OAIDCtoMARCXML.xsl* crosswalk, it became evident that further customization would be necessary to make use of all the data available from the harvest of Penn State's ETD website, particularly regarding new mappings of degree type, graduate degree program, and access restrictions.[27] After ten separate revisions of the original crosswalk, members of the Digital Access Team working with the Authority Control Librarian of the Cataloging and Metadata Services Department reviewed samples of several hundred MARC records. Through each iteration, feedback was provided, errors were noted, and corrections and modifications were made to the crosswalk until its output met the department's quality standards.

One of the authors had prior programming experience, but no experience in XSLT coding. An online tutorial and a standard reference book were used to acquire a basic understanding of XSLT before making changings to the XSLT crosswalk.[28] Further information used in helping to debug the crosswalk was obtained by searching forums at Stack Overflow.[29] In total, 92.5 hours spread out over a seven-month timeframe were used to learn XSLT, code the crosswalk, debugging, testing, getting feedback, and writing documentation. While this amount of time is considerable, the initial investment paid dividends almost immediately, as the time required to process a semester's worth of ETDs plummeted from 100–160 hours to fewer than 8 hours. Return on investment (92.5 hours) occurred as soon as the new procedure was implemented in addition to paying dividends: 10–70 hours of newly available staff time.

The first customizations made on the XSLT crosswalk handled local non-standard assignments to the DC elements *coverage*, *relation*, and *rights* that are discussed in the previous section. *Coverage* contained the name of the graduate degree program, such as Architecture, Aerospace Engineering, and Kinesiology. The original *OAIDCtoMARCXML.xsl* crosswalk mapped this to the MARC 500 field. This was changed to MARC 699, a local subject access field specifically for collating theses by graduate degree program in The CAT. In initial tests, the MARC records output by the harvest showed MARC 699 positioned between MARC 300 and MARC 520. To correct this, the code for mapping MARC 699 in the XSLT crosswalk was moved between mappings of MARC 653 and MARC 700.

Similar changes were made to other DC elements. The DC element *relation* contained the degree type, such as PhD or MS. The original *OAIDCtoMARCXML.xsl* crosswalk had mapped it to the MARC 787 field. Because degree types belong in MARC 502, the code in the crosswalk was changed to map it to that field. The DC element *rights* contained access restrictions. There were three possible values for this DC element: "Open Access," "Restricted," and "Restricted (Penn State Only)." In the original crosswalk, *rights* was mapped to MARC 540 (Terms Governing Use and Reproduction Note). This was changed to MARC 506 (Restriction on Access Note). The DC element *subject* was remapped from MARC 690 to MARC 653, since this DC element only contained author-supplied keywords. These three remaps were also re-positioned in the crosswalk so that their output displayed in the correct positions within a MARC record.

The second stage of customization involved adding MARC fields to the crosswalk that contained constant data appearing in every thesis MARC record. Examples included the MARC fields 006, 007, 008, 040, 260/264 (Production, Publication, Manufacture and Copyright Notice), 300 (Physical Description), and 538 (System Details Note). The MARC 008 was a special case: positions 00–05 required a computer-generated, six-character numeric string indicating the date the record was created in the format yymmdd. A function that could retrieve the current date was required. XSLT uses a language called XML Path Language, or XPath, that addresses parts of an XML document and performs calculations on it.[30] XPath provides several functions that XSLT can use. One of these is a function that retrieves the current date, *current-date()*, but because this function is from XPath

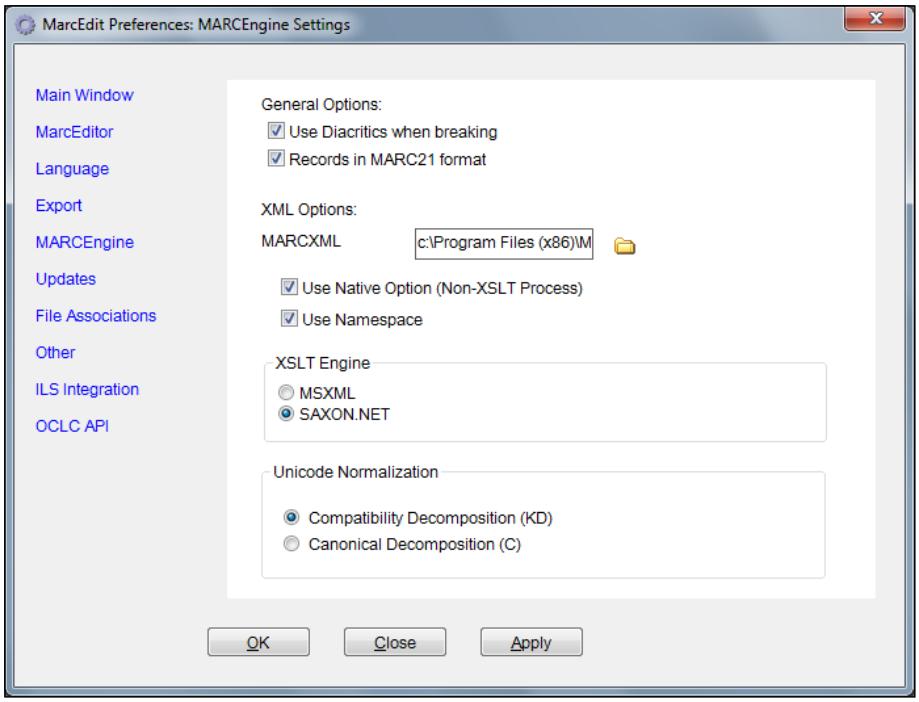

Figure 2. MARCEngine Settings in MarcEdit

2.0 (with the current crosswalk in XSLT 1.0), the XSLT Engine in MarcEdit needed to be set to SAXON.NET, an XSLT processor designed to run using the Microsoft .NET Framework, in the MARCEngine section of MarcEdit's preferences. By default, MarcEdit is set to the MSXML XSLT Engine, which does not support XPath 2.0 functions. Changing the XSLT Engine allowed us to add version 2.0 functions to a version 1.0 XSLT crosswalk without having to upgrade the entire crosswalk to version 2.0. This saved coding time, but in the future it may be desirable to convert the entire crosswalk to XSLT 2.0. The output of this function yielded the date in the format yyyy-mm-dd-hh:mm. To convert this date into the format needed for MARC 008, the output was concatenated using three separate *substring()* functions together.[31] Figure 2 shows the MARCEngine settings in MarcEdit used for the customized crosswalk.

In addition to retrieving the record creation date, the publication date is also required for MARC 008 positions 7–10 (Date1) and MARC 260/264 subfield $c. This was obtained during the harvest from DC element *date*. The date was in yyyy-mm-dd format and with the use of a *substring()* function, the first four characters were mapped to all of these MARC21 positions.

During the early stages of testing, harvested data was output into MARC records using AACR2. The department began the transition to RDA in early 2013, but thesis cataloging had not yet made the transition at the time of testing. This was an opportune time to convert the XSLT crosswalk to output RDA-compliant MARC data. Following Program for Cooperative Cataloging (PCC) guidelines for RDA records, MARC 260 was changed to MARC 264.[32] The General Material Designation (GMD) in MARC 245 subfield $h was replaced with three new MARC fields: 336 (Content Type), 337 (Media Type), and 338 (Carrier Type). Instead of using MARC 502 subfield $a (Dissertation Note), dissertation information was parsed into separate subfields: $b (Degree Type), $c (Name of Granting Institution), and $d (Year Degree Granted). Relationship designators were added to MARC 100 and 700 in subfield $e. In other areas of the MARC record, abbreviations were spelled out, such as converting "Pa." to "Pennsylvania" in MARC 264 subfield $a (Place of Production, Publication, Distribution, Manufacture).

The next area customized was the display of degree type in MARC 502 $b. ETD authors submitted this information via a dropdown box, but the format in which this information was stored in DC element *relation* did not coincide with the punctuation conventions currently used in 502 $b, for example, Ph.D. was stored as "PHD" and M.Agr. as "M AGR." XSLT provides a method for expressing multiple conditional tests, by using an <*xsl:choose*> element in conjunction with multiple <*xsl:when*> elements.[33] An <*xsl:otherwise*> element is used in conjunction with <*xsl:choose*> when none of the <*xsl:when*> elements matches the conditions being tested. To use this method, the degree type harvested from DC element *relation* was stored in an XSLT variable called *degree*. Ten different <*xsl:when*> tests were performed on the value in *degree* to see if it matched one of ten different degree types offered by Penn State. If it matched one of them (i.e., PHD), the corrected form (i.e., Ph.D.) was stored in another variable called *degree_output*. The value in *degree_output* was mapped to 502 $b. If none of the <*xsl:when*> tests resulted in a match, then the value in <*xsl:otherwise*> was used. The term "Unknown" was assigned for this case. This will be helpful for detecting any future new degree types. Each time a harvest is performed, a visual scan of the records is sufficient to catch these for manual correction and future updating of the crosswalk. Figure 3 shows the XSLT coding for mapping the 502 field.

A similar <*xsl:choose*> structure was created for a single subfield in a local MARC 949 field. At Penn State, the 949 field is used to create holdings information for each record

during batchload into the catalog. The field contains nine subfields of which eight are constant data, set by local policy and coded directly into the XSLT crosswalk:

- $a (Call Number) = Electronic thesis
- $w (Class Scheme) = ASIS
- $m (Library) = ONLINE
- $k (Current Location) = ONLINE
- $l (Home Location) = ONLINE
- $o (Notes) = no value assigned
- $r (Circulate Flag) = Y
- $s (Permanent Flag) = Y

The ninth, subfield $t contains the item type. For ETDs, only two values are valid: THESIS-D for doctoral dissertations and THESIS-M for master's theses. These values are used in The CAT as limits for searching on doctoral dissertations or master's theses. The <*xsl:choose*> coding for this was almost the same as that for mapping to 502 $b, except it mapped an item type based on the value found in DC element *relation*. For example, "PHD" maps to THESIS-D and "MS" maps to THESIS-M. The <*xsl:otherwise*> value was set to THESIS-M because there were larger numbers of master's theses than doctoral dissertations during testing. This will prevent this subfield from being blank and causing a batchload to fail. During a visual scan following a harvest, any instances of "Unknown" found in 502 $b requires that the cataloger check and correct 949 $t. Figure 4 shows the XSLT coding for using the degree type to determine the item type in Penn State's local 949 field.

```
<xsl:variable name="degree" select="dc:relation"/>
- <xsl:variable name="degree_output">
  - <xsl:choose>
      <xsl:when test="$degree='PHD'">Ph.D.</xsl:when>
      <xsl:when test="$degree='DED'">D.Ed.</xsl:when>
      <xsl:when test="$degree='DMA'">D.M.A.</xsl:when>
      <xsl:when test="$degree='MS'">M.S.</xsl:when>
      <xsl:when test="$degree='M AGR'">M.Agr.</xsl:when>
      <xsl:when test="$degree='MArch'">M.Arch.</xsl:when>
      <xsl:when test="$degree='MA'">M.A.</xsl:when>
      <xsl:when test="$degree='M Ed'">M.Ed.</xsl:when>
      <xsl:when test="$degree='ME'">M.Eng.</xsl:when>
      <xsl:when test="$degree='MLA'">M.L.A.</xsl:when>
      <xsl:otherwise>Unknown</xsl:otherwise>
    </xsl:choose>
</xsl:variable>
- <datafield tag="502" ind2=" " ind1=" ">
  - <subfield code="b">
      <xsl:value-of select="$degree_output"/>
    </subfield>
    <subfield code="c">Pennsylvania State University</subfield>
  - <subfield code="d">
      <xsl:value-of select="concat(substring(dc:date,1,4), '.')"/>
    </subfield>
</datafield>
```

Figure 3. XSLT Coding for Mapping Degree Type in MARC 502 Field

Coding was added to the XSLT crosswalk to handle initial articles in thesis titles. Because the majority of Penn State theses are written in English, the crosswalk handles only the initial articles "a," "an," and "the." Respectively MARC 245 indicator position two is set to 2, 3, and 4. In all other cases, it is set to 0.

Another challenge was determining where a title ends and a subtitle begins. Sharretts, Shieh, and French noted that they considered anything following a colon as a subtitle.[34] Penn State took a similar approach, but expanded it to include the space following the colon. This decision was made in anticipation of unusual usage of colons in acronyms or for artistic or typographical effects. Our samples showed that the space following the colon was used in all cases and future testing will determine whether more elaborate coding is warranted.

Author names were already stored in inverted order on Penn State's ETD website and did not contain fuller forms or birth dates. Consequently, the original code in the *OAIDCtoMARCXML.xsl* crosswalk was simplified by removing the "persname_template," an XSLT template designed to construct the MARC 100 and 700 fields for names with fuller forms and/or birth dates. Functionality for mapping additional authors to MARC 700 was retained even though co-authors were not found among any of the samples tested.

Unlike author names, thesis advisor and committee member names were stored in the DC element *contributor* in direct order. The form that ETD authors used to submit their thesis advisors and committee members is in free format, though there are separate areas for the advisors and committee members. In addition to the name, DC element *contributor* contains the role the individual played following the name and separated by a semicolon and space character. Roles include Thesis Advisor, Dissertation Advisor, Committee Chair, and Committee Member. Examples include "Jane Doe; Thesis Advisor" or "John Doe; Committee Member." There can be multiple thesis advisors and multiple committee members for each thesis. When a thesis advisor's name is not present, the committee chair is assumed to be the thesis advisor. Adding to the complexity, names as entered by the ETD author sometimes include prefixes (Dr., Professor), suffixes (Jr., III), and the degree of the advisor or committee member (PhD, DEd). In some rare cases, several members' names were entered consecutively in the same field. The

goal was to get all thesis advisors associated with a thesis mapped to MARC 700 fields with their names in indirect order. This was a particularly challenging and complicated coding task.

An Open Archives Initiative harvest of 773 theses was used as a sample to determine the variations found in the DC *contributor* element. Each variation was noted and an algorithm was developed to address the most common forms and some of the more prevalent problematic forms. While processing the *contributor* element, any unusual findings were mapped to the MARC 720 (Added Entry—Uncontrolled Name) for evaluation after the harvest. As a backup, anything the algorithm missed will be detected by our authority control vendor and reported as errors that can be cleaned up later. In future harvests, the algorithm may need to be adjusted to address new issues that may arise.

```
<xsl:variable name="degree" select="dc:relation"/>
- <xsl:variable name="itemtype">
  - <xsl:choose>
      <xsl:when test="$degree='PHD'">THESIS-D</xsl:when>
      <xsl:when test="$degree='DED'">THESIS-D</xsl:when>
      <xsl:when test="$degree='DMA'">THESIS-D</xsl:when>
      <xsl:when test="$degree='MS'">THESIS-M</xsl:when>
      <xsl:when test="$degree='M AGR'">THESIS-M</xsl:when>
      <xsl:when test="$degree='MArch'">THESIS-M</xsl:when>
      <xsl:when test="$degree='MA'">THESIS-M</xsl:when>
      <xsl:when test="$degree='M Ed'">THESIS-M</xsl:when>
      <xsl:when test="$degree='ME'">THESIS-M</xsl:when>
      <xsl:when test="$degree='MLA'">THESIS-M</xsl:when>
      <xsl:otherwise>THESIS-M</xsl:otherwise>
    </xsl:choose>
</xsl:variable>
- <datafield tag="949" ind2=" " ind1=" ">
    <subfield code="a">Electronic thesis</subfield>
    <subfield code="w">ASIS</subfield>
    <subfield code="m">ONLINE</subfield>
    <subfield code="k">ONLINE</subfield>
    <subfield code="l">ONLINE</subfield>
    <subfield code="r">Y</subfield>
    <subfield code="s">Y</subfield>
  - <subfield code="t">
      <xsl:value-of select="$itemtype"/>
    </subfield>
</datafield>
```

Figure 4. XSLT Coding for Setting Item Type Based on Degree Type

A simplified version of the algorithm to convert direct-order personal names as harvested into name added entries in the MARC bibliographic record in inverted order (i.e., Last Name, First Name, other data):

1. Gather the roles of the first nine contributors (an arbitrary value intended to exceed the typical number of possible individuals).
2. Separate the name from the role using the position of the semicolon.
3. If the role is "Committee Chair" and no other role contains the term "Advisor," then set the role for that individual as "Thesis Advisor."
4. Because the name in DC element *contributor* is in direct order, assume the presence of a comma to mean the name contains a suffix or degree following it. Separate the data following the comma from the name. An example would be Martha Evans, PhD.
5. If the role is "Dissertation Advisor" or "Thesis Advisor," continue to the next step. Otherwise, ignore this name, exit the algorithm, and then start the whole process over with the next name.
6. Remove any titles from the beginning of the name (Dr., Prof., etc.).
7. Tokenize the name (i.e., split the name into individual elements).
8. Check the last token. If it contains a suffix (Jr., Sr., III, etc.), choose the second to last token as the surname. Otherwise, the last token is the surname. This check for suffixes is because sometimes they appear before the comma.
9. Output the surname into a MARC 700 subfield $a, followed by a comma, and then the rest of the tokens preceding the last token (first and middle names or initials).
10. If the name contains a suffix (Jr., Sr., III, etc.), then output the suffix into MARC 700 subfield $c.
11. Output "thesis advisor" into MARC 700 subfield $e.
12. Discard any titles or degree information following the comma (PhD, MS, Prof., etc.). After discarding, if any remaining data are found, output into MARC 720.

This algorithm may not appear to follow a logical pattern. The apparent illogic is due to having to write code within the limitations of variable assignments in XSLT. Unlike many traditional programming languages, XSLT does not allow a variable's value to be modified once it has been set.[35]

In testing the customized XSLT crosswalk, there was concern about how to handle unusual characters that might encode or display incorrectly in Penn State's online catalog

or in OCLC WorldCat. This occurred when a thesis was found to contain a Unicode line separator character, which caused half of a summary note (MARC 520) to appear at the end of the record in MarcEdit's MarcEditor. It became apparent that thesis authors often copied and pasted much of their information from whatever word processing software they used into Penn State's thesis submission forms. This practice introduced a large number of characters not generally compatible with online library catalogs. Penn State's Digital Library Technologies staff partially addressed the issue by applying a filter to strip control characters from the OAI-PMH feed. During additional troubleshooting, a sample harvest was imported into a local save file in the OCLC Connexion client. A considerable number of these records did not validate because of incompatible characters.

A two-pronged approach was used to address these characters. First, a script was written using AutoIt that reads in a MarcEdit .mrk file and writes the record numbers and incompatible characters found in those records into a spreadsheet.[36] Second, after reviewing the resulting spreadsheet, an XSLT template was developed for the crosswalk to convert all of the incompatible characters found in the 773-record sample harvest noted above.

The first process the template performed was a Unicode normalization using the Normalization Form Compatibility Decomposition (NKFD).[37] This converted single characters (a letter and a diacritic as a single character) into their decomposed forms of a letter and a combining mark. It also separated ligatures (such as "fi") into two separate characters. There are four different Unicode normalization forms, and through testing, NKFD produced the best results. For this to work, the XSLT Engine settings in MarcEdit required that Unicode Normalization be set to Compatibility Decomposition (KD).

The remainder of the template converts individual non-compatible characters into their compatible equivalents or as bracketed interpolations. These included both lowercase and uppercase Greek letters used as mathematical variables, right and left quotes and double quotes, a variety of dash and hyphen symbols, a large number of characters found in the Unicode Private Use Area for Microsoft symbol fonts, and other assorted mathematical symbols (such as the infinity symbol).

Because there is no way to predict what kinds of non-compatible characters thesis authors might include in their metadata, and developing a template to handle thousands of such characters is time-consuming, the AutoIt script used to detect them will be part of the workflow for future harvests. Henceforth, any non-compatible characters found will be manually corrected before loading into Penn State's online catalog. During future reviews, non-compatible characters that appear in large numbers may warrant additions to the crosswalk's template.

Throughout the development of this customized template, the Digital Access Team invested a significant amount of effort to minimize the amount of work needed in future harvests. The team expects that further tweaks will be necessary; the decision to implement a given enhancement will be based on whether the time required to implement the enhancement will save time or resources in the future.

New Workflow, 2014–

As noted above, the old workflow was largely manual. The new workflow was designed to free up a copy cataloger's time by leveraging the power of harvesting author-supplied data and batch loading records into the catalog. The time required to develop and test the process was not insignificant, but the immediate and long-term savings in time and gains in efficiency and quality warranted the decision to invest resources up front. The new process is outlined below.

After receiving the .mrk file of ETD metadata, the thesis cataloger begins quality assurance procedures. Using MarcEdit's Export Tab Delimited function, the cataloger exports a set of fields into a tab-delimited text file which is then opened in Excel to allow for sorting by fields. The included MARC fields are 001 (control number), 100$a (author name), 245 (title), 502 (degree type), 506 (access note), 699 (academic program), 700 (advisor), 856$u (URL), 720 (used for advisor fields that could not be properly parsed), 949 (holdings note, which generates call number and item information), and 520$a (summary).

The cataloger first sorts by author's name (MARC 100) and compares the file with a list of ETDs provided by the Graduate School to ensure that all ETDs were included in the extract. It may be necessary to either manually catalog ETDs that are missing from the extract, or to delete ETDs from the .mrk file that are not on the Graduate School's list for that semester. Some discrepancy in names may be caused by name changes, different parsing of compound surnames, or misspellings. The cataloger then compares the 502 field (degree type) to ensure that it matches the item type contained in the MARC 949, either THESIS-D for doctoral degrees (PhD, DEd, etc.) or THESIS-M for master's degrees (MS, MA, MEng, etc.). The cataloger checks for the presence of any 720 fields, which indicate that manual intervention is required on the 700 advisor note, and resolves as necessary. Next, the cataloger scans the 506 field and makes a list of all "Restricted" files, to shadow these records in The CAT after load. In SirsiDynix parlance, shadowing a record leaves it intact in the catalog but removes it from public view.

The cataloger checks for any ETDs that require assignment of LCSH based on departmental standards. These records are referred to original catalogers after the file has been loaded into The CAT. The cataloger uses the Find function in Excel on the column containing MARC 245 data

to find references to Pennsylvania or Penn State in ETD titles. The cataloger sorts by Graduate Degree Program (in the 699 field) and manually scans the titles of all ETDs in departments (e.g., humanities, arts, languages) most likely to be associated with such data to check for references to authors and works. At this stage, the cataloger also scans for ETD titles in non-English languages, and corrects the language fixed field for those records.

Next, the cataloger uses the AutoIt script, which checks all fields in the .mrk file for text characters not compliant with OCLC load requirements. After manually correcting these in the .mrk file, the ETD cataloger compiles the file into an .mrc file that is ready to load into SirsiDynix's staff module WorkFlows. When the load is complete, the cataloger shadows all records for "Restricted" ETDs to hide them from public view, and refers any records requiring subject headings to an original cataloger. Access restrictions last up to two years, after which the cataloger receives notification from the Graduate School and makes the records for shadowed ETDs visible to the public. For examples of thesis MARC records from 2002 (when many theses were still produced in print), 2004–14 (when ETDs were cataloged by hand), and from 2014– (after the adoption of the new workflow), see the Appendix.

Conclusion

Ensuring discovery of and access to materials in low-barrier self-deposit services, such as ETD databases, requires an enormous investment of time and resources when approached with a traditional cataloging mindset, i.e., cataloging items one-by-one. By leveraging metadata supplied by authors at the time of deposit, OAI-PMH harvests, and the transformations of data possible with XSLT, the authors devised tools and a workflow that greatly improved the efficiency of the cataloging process with minimal impact on metadata quality. Development and testing of the new procedure required a considerable investment of time, but with the scripts now in place and a redesigned workflow, a procedure that previously required months of staff time annually now takes hours. As cataloging and metadata departments are being asked to provide new services while still keeping up with traditional workflows, it is imperative to make every effort to streamline procedures that can be simplified. Cataloging ETDs is one such procedure. By extension, variants of the tools and processes described above could be applied to similar cases, such as institutional repositories or, in fact, any database in which metadata resides and is harvestable via OAI-PMH.[38]

References and Notes

1. Penn State University Libraries, Cataloging and Metadata Services, Digital Access Team, "Digital Access Team" (University Park, PA: Penn State University Libraries, 2015), www.libraries.psu.edu/psul/cataloging/digital.html.
2. Penn State University Libraries, "The CAT" (University Park, PA: Penn State University Libraries, 2014), http://cat.libraries.psu.edu.
3. Pennsylvania State University, "ETD, Electronic Theses and Dissertations" (University Park, PA: Pennsylvania State University, 2011), https://etda.libraries.psu.edu.
4. Open Archives Initiative, "Open Archives Initiative Protocol for Metadata Harvesting" (n.p.: Open Archives Initiative, n.d.), accessed November 23, 2015, www.openarchives.org/pmh/.
5. MarcEdit Development, "MarcEdit—Your complete free MARC editing utility" (n.p.: MarcEdit Development, 2013), http://marcedit.reeset.net.
6. American Library Association, "RDA Toolkit" (Chicago: American Library Association, 2010), www.rdatoolkit.org.
7. Cristina W. Sharretts, Jackie Shieh, and James C. French, "Electronic Theses and Dissertations at the University of Virginia," in Proceedings of the *ASIS Annual Meeting*, 240–55 (Medford, NJ: Information Today, 1999).
8. Brian E. Surratt and Dustin Hill, "ETD2MARC: A Semi-Automated Workflow for Cataloging Electronic Theses and Dissertations," *Library Collections, Acquisitions, & Technical Services* 28, no. 2 (2004): 205–23, http://dx.doi.org/10.1016/j.lcats.2004.02.014; Galen Charlton, "MARC-Record-2.0.6" (n.p.: CPAN, 2013), http://search.cpan.org/~gmcharlt/MARC-Record-2.0.6/.
9. Sharon Reeves, "User-Generated Metadata for ETDs: Added Value for Libraries," in Proceedings of the Tenth *International Symposium on Electronic Theses and Dissertations* (Uppsala: Uppsala University Library, 2007), http://epc.ub.uu.se/ETD2007/files/papers/paper-40.pdf.
10. Sevim McCutcheon et al., "Morphing Metadata: Maximizing Access to Electronic Theses and Dissertations," *Library Hi Tech* 26, no. 1 (2008): 41–57, http://dx.doi.org/10.1108/07378830810857799.
11. "MARC/Perl" (n.p. : SourceForge, n.d.), accessed November 23, 2015, http://marcpm.sourceforge.net.
12. Terry Reese, "Automated Metadata Harvesting: Low-Barrier MARC Record Generation from OAI-PMH Repository Stores Using MarcEdit," *Library Resources & Technical Services* 53, no. 2 (April 2009): 121–34, www.ala.org/alcts/sites/ala.org.alcts/files/content/resources/lrts/archive/53n2.pdf.
13. Michael Boock and Sue Kunda, "Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries," *Cataloging & Classification Quarterly* 47, no. 3–4 (2009):

297–308, http://dx.doi.org/10.1080/01639370902737323.
14. Sai Deng and Terry Reese, "Customized Mapping and Metadata Transfer from DSpace to OCLC to Improve ETD Work Flow," *New Library World* 110, no. 5–6 (2009): 249–64, http://dx.doi.org/10.1108/03074800910954271.
15. Merry Bower, Martin Courtois, and Michelle Turvey-Welch, "ETDs Transformed: Maximizing Cataloging Efficiencies and Open Access," in *International Symposium on Electronic Theses and Dissertations* (ETD 2009) (Pittsburgh: University of Pittsburgh, 2009), http://docs.ndltd.org/dspace/handle/2340/1129.
16. Maureen P. Walsh, "Metadata Repurposing Using XSLT," *More Technology For the Rest of Us: A Second Primer on Computing for the Non-IT Librarian* (Santa Barbara, CA: Libraries Unlimited, 2010), 125–39.
17. Shawn Averkamp and Joanna Lee, "Repurposing ProQuest Metadata for Batch Ingesting ETDs into an Institutional Repository," *Code4Lib Journal* 7 (2009), http://journal.code4lib.org/articles/1647.
18. Cedar C. Middleton, Jason W. Dean, and Mary A. Gilbertson, "A Process for the Original Cataloging of Theses and Dissertations," *Cataloging & Classification Quarterly* 53, no. 2 (2015): 234–46, http://dx.doi.org/10.1080/01639374.2014.971997.
19. Sevim McCutcheon, "Basic, Fuller, Fullest: Treatment Options for Electronic Theses and Dissertations," *Library Collections, Acquisitions & Technical Services* 35 (2011): 64–68, http://dx.doi.org/10.1016/j.lcats.2011.03.019.
20. Virginia Tech, "etds@vt" (Blacksburg: Virginia Tech, 2016), https://theses.lib.vt.edu/ETD-db/index.shtml.
21. Django Software Foundation, "Django" (n.p.: Django Software Foundation, 2016), accessed February 5, 2016, https://www.djangoproject.com; Oracle Corporation, "MySQL" (n.p.: Oracle Corporation, 2016), www.mysql.com.
22. Insight Software Solutions, "Macro Express: The Windows Automation Tool" (Kaysville, UT: Insight Software Solutions, 2015), www.macros.com.
23. AutoIt Consulting, "AutoIt: Automation and Scripting Language" (n.p.: AutoIt Consulting, 2014), www.autoitscript.com/site/autoit/; Ken Robinson, "PSU Thesis PDF Title to WorkFlows Conversion" (University Park: Penn State University Libraries, 2015), www.libraries.psu.edu/psul/cataloging/techresources/macros/autoit_docs/thesis_pdf.html.
24. Ken Robinson, "653 Creation from PSU Thesis Keywords" (University Park: Penn State University Libraries, 2015), www.libraries.psu.edu/psul/cataloging/techresources/macros/autoit_docs/create653.html.
25. Penn State, The Graduate School, "Thesis and Dissertation Guide: Requirements and Guidelines for the Preparation of Master's Theses and Doctoral Dissertations" (University Park, PA: Office of Theses and Dissertations, n.d.), accessed November 23, 2015, www.gradschool.psu.edu/current-students/etd/thesisdissertationguidepdf/.
26. Pennsylvania State University, "EHT, Electronic Honors Theses" (University Park, PA: Pennsylvania State University, 2011), https://honors.libraries.psu.edu.
27. The crosswalk and a description of how it works is available for download. See Ken Robinson, "Penn State eTD Dublin Core-to-MARCXML Crosswalk" (University Park: Penn State ScholarSphere, 2015), https://scholarsphere.psu.edu/collections/x346dj68d.
28. W3Schools, "XSLT Tutorial" (n.p.: W3Schools, 2015), accessed November 23, 2015, www.w3schools.com/xsl/; Jeni Tennison, "Beginning XSLT 2.0: From Novice to Professional" (Berkeley, CA: Apress, 2005).
29. "Stack Overflow" (n.p.: Stack Exchange, 2015), http://stackoverflow.com.
30. James Clark and Steve DeRose, "XML Path Language (XPath), Version 1.0" (n.p.: W3C, 1999), http://www.w3.org/TR/xpath/.
31. W3Schools, "XSLT, XPath, and XQuery Functions," *XSLT Tutorial* (n.p.: W3Schools, 2015), www.w3schools.com/xsl/xsl_functions.asp.
32. Program for Cooperative Cataloging, "RDA PCC Proposed Guidelines and Standards" (Washington, DC: Program for Cooperative Cataloging, 2015), www.loc.gov/aba/pcc/rda/RDA%20PCC%20Proposed%20Guidelines%20and%20Standards.html.
33. W3Schools, "XSLT <xsl:choose> Element," *XSLT Tutorial* (n.p.: W3Schools, 2015), www.w3schools.com/xsl/xsl_choose.asp.
34. Sharretts, "Electronic Theses," 251.
35. W3Schools, "XSLT <xsl:variable> Element," *XSLT Tutorial* (n.p.: W3Schools, 2015), www.w3schools.com/xsl/el_variable.asp.
36. The program code for this AutoIt script is available for download. See Ken Robinson, "AutoIt Script for Identifying OCLC Non-Compatible Characters in a MarcEdit .mrk File" (University Park: Penn State ScholarSphere, 2015), https://scholarsphere.psu.edu/files/x346dv29k.
37. Mark Davis and Ken Whistler, eds., "Unicode Standard Annex #15: Unicode Normalization Forms (Mountain View, CA: Unicode Consortium, 2014), http://unicode.org/reports/tr15.
38. It should be noted that subsequent to the adoption of the new ETD workflow, the Digital Access Team devised similar workflows, based on OAI-PMH and batch loading records, to catalog items available through the NASA Scientific and Technical Information Program (the NASA Technical Reports Server, www.sti.nasa.gov) and selected collections with good metadata in the Internet Archive.

Appendix

A typical print thesis MARC record from 2002:

LDR 00638 tm a2200169 a 4500
008 030902s2002\\\\pau\\\\\\\\\\000\0\eng\d
040 __ $a PSt $c PSt
100 1_ $a Halperin, Gregory S.
245 10 $a Therapist personality and tactics of interpersonal influence /$c by Gregory S. Halperin.
260 __ $c 2002.
300 __ $a 83 leaves.
502 __ $a Thesis (Ph.D.)--Pennsylvania State University, 2002.
533 __ $a Microfilm (positive). $e 1 reel 35 mm. $c (University Microfilms 30-64931)
699 __ $a Psychology.

A typical ETD MARC record from the Old Workflow period, 2004–14:

LDR 01428cam a22003137 4500
006 m\\\\\\\d\\\\\\
007 cr\ln\\\\\\\n
008 121205s2012\\\\pau\\\\\\\\\\000\0\eng\u
040 __ $a PSt $c PSt
100 1_ $a Shipley, Peggy Zoe.
245 10 $a Life patterns of family caregivers of patients with amyotrophic lateral sclerosis $h [electronic resource] / $c by Peggy Zoe Shipley.
246 30 $a Family caregivers of patients with amyotrophic lateral sclerosis
260 __ $a [University Park, Pa.] : $b Pennsylvania State University, $c 2012.
300 __ $a 1 electronic document (443 p.)
500 __ $a Mode of access: World Wide Web.
500 __ $a Thesis advisor: Janice Penrod.
502 __ $a Thesis (Ph.D.)--Pennsylvania State University, 2012.
533 __ $aMicrofilm (positive). $e 1 reel ; 35 mm. $c (University Microfilms 35-34763)
538 __ $a The full text of the dissertation is available as a Adobe Acrobat .pdf file ; Adobe Acrobat Reader required to view the file.
653 0_ $a Family caregiver; health as expanding consciousness; ALS; amyotrophic lateral sclerosis; life pattern
699 __ $a Nursing.
700 1_ $a Penrod, Janice L., $e thesis advisor.
856 40 $u https://etda.libraries.psu.edu/paper/15131/

A typical ETD MARC record harvested using the New Workflow, 2014–:
LDR 03310nam a22003737i 4500
006 m\\\\\|\\d\\\\\\
007 cr\ln\\\\\\n\\
008 141027s2014\\\\pau\\\\\sm\\\\000\0\eng\d
040 __ $a PSt $b eng $e rda $c PSt
042 __ $a dc
100 1_ $aTownsend, Andrew, $e author.
245 10 $aRevisiting Mesopredator Release : $b Carnivore Dynamics Along A Gradient of Landscape Disturbance.
264 _1 $a [University Park, Pennsylvania] : $b Pennsylvania State University, $c 2014.
300 __ $a 1 electronic document.
336 __ $a text $b txt $2 rdacontent
337 __ $a computer $b c $2 rdamedia
338 __ $a online resource $b cr $2 rdacarrier
502 __ $b M.S. $c Pennsylvania State University $d 2014.
506 __ $a Open Access.
520 __ $a Human induced habitat loss and predator persecution caused severe declines in apex carnivores throughout the North American continent. Removal of apex predators allowed smaller, lower rank predators from the Order Carnivora to become prominent. These "mesopredators" flourished, destabilizing ecosystems by driving many prey species toward extinction. However, some suggest that mesopredators still benefit from contemporary vegetation changes and fragmentation by thriving in disturbed areas. Many worry the versatility of these mesopredators could further threaten their prey species by leading to increased predation in anthropogenically-disturbed areas. This study seasonally sampled predator distributions along land cover gradients in forested, riparian corridors in Appalachia to identify whether landscape modification results in changes in carnivore community structure in the region. The study area consisted of randomly generated sites along streams in central Pennsylvania. I gathered data from camera traps and field surveys to catalogue the spatial ecology of mesopredators. I analyzed these data with landscape metrics to test the hypothesis that as forest contiguity decreases, both the abundance and richness of the predator community increases, possibly adding pressure on vulnerable prey populations. Through the analysis of these habitat metrics and carnivore occurrence data, this study found that carnivore species richness and abundance do generally increase with human disturbance in rural settings. However, this pattern is not due to the behavior of every species as many mesopredators are present across these rural landscapes and exhibit different responses to disturbance. Nevertheless, a few important generalists, namely the canids and raccoons, do show preferences toward more human disturbed areas and thus, are most accountable for this observed pattern.
538 __ $a The full text of the dissertation is available as an Adobe Acrobat .pdf file ; Adobe Acrobat Reader required

to view the file.
653 _0 $a spatial ecology
653 _0 $a mesopredator
653 _0 $a predation
653 _0 $a landscape gradients
653 _0 $a riparian corridors
653 _0 $a Appalachia
653 _0 $a ecological cascades
699 __ $a Geography
700 1_ $a Brooks, Robert P., $e thesis advisor.
700 1_ $a Bishop, Joseph A., $e thesis advisor.
700 1_ $a Serfass, Thomas L., $e thesis advisor.
856 40 $u https://etda.libraries.psu.edu/paper/22618 $z Connect to this object online.
949 __ $a Electronic thesis $w ASIS $m ONLINE $k ONLINE $l ONLINE $r Y $s Y $t THESIS-M