Notes on Operations

A Case Study of ETD Metadata Remediation at the University of Houston Libraries

Santi Thompson, Xiping Liu, Albert Duran, and Anne M. Washington

This paper provides a case study on remediating electronic theses and dissertations (ETD) metadata at the University of Houston Libraries. The authors provide an overview of the team's efforts to revise existing ETD metadata in its institutional repository as part of their commitment to aligning ETD records with the Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations, Version 2.0 (TDL guidelines, version 2). The paper reviews the existing literature on metadata quality and ETD metadata practices, noting how their case study adds one of the first documented cases of ETD metadata remediation. The metadata upgrade process is described, with close attention to the tools and workflows developed to complete the remediation. The authors conclude the paper with a discussion of lessons learned, the project's limitations, future plans, and the emerging needs of metadata remediation work.

Over the last two decades, institutions have increasingly accepted electronic theses and dissertations (ETDs) as part of a student's graduation requirements. Not surprisingly, the proliferation of these documents have prompted libraries and other stakeholder groups to confront policy and workflow issues addressing the curation of digital objects from acquisition to preservation, including submission protocols, document embargo options, and promoting access. In the process of confronting these issues, librarians and information professionals have developed common and best practices regarding how ETDs are described, often focusing on the benefits and limitations of certain metadata schema, the number of types of metadata fields necessary to adequately describe a work, and the challenges incurred through accepting author-generated metadata.

While building on the previous work of ETD metadata research, this paper provides a case study for another aspect of ETD description: metadata remediation. For the purposes of this paper, the authors define metadata remediation as the process of evaluating previously generated metadata, either user- or library-created, and refining it based on shifting institutional practices and updated metadata standards. While the literature has a growing body of work dedicated to metadata creation and quality review, it lacks documented cases of ETD metadata remediation. As a result, there are few examples of shared lessons to consider when undertaking a remediation project or common approaches to begin drafting best practices.

The authors will begin by providing an overview of the University of Houston (UH) Libraries' efforts to revise existing ETD metadata in its institutional repository as part of their commitment to align ETD records with the TDL guidelines, version 2.[1] After a brief background and history of UH's ETD program, the authors review the existing literature on metadata quality and ETD metadata practices, noting how their case study adds an additional documented case of metadata remediation. They then describe their metadata upgrade process, with close attention to the tools and workflows developed to complete

Santi Thompson (sathomp3@central.uh.edu) is the Head of Digital Research Services, University Libraries, University of Houston. **Xiping Liu** (xliu54@central.uh.edu) is a Resource Description Librarian, University Libraries, University of Houston. **Albert Duran** (Albert.Duran@houstontx.gov) is a Metadata Librarian at the Houston Public Library, Houston, Texas. **Anne M. Washington** (awashin8@central.uh.edu) is Metadata Services Coordinator, University Libraries, University of Houston.

Manuscript submitted February 2, 2018; returned to authors for revision March 30, 2018; revised manuscript submitted July 24, 2018; accepted for publication September 25, 2018.

the remediation. The paper concludes with a discussion of lessons learned, project limitations, future plans, and the emerging needs of metadata remediation work. While this use case is especially suited for smaller collections (approximately one thousand records), the workflow takes advantage of commonly known tools and simple steps, making it accessible and extensible for other use cases.

Overview of UH's ETD Program

UH is a Carnegie-designated tier 1, doctoral granting research university with over 40,000 enrolled students, 2,500 faculty members, and nearly 200 graduate degree programs. In 2009, the Faculty Senate's Graduate and Professional Studies Committee approved a new policy requiring all graduate programs producing a thesis or dissertation to migrate to electronic format by summer 2014. UH colleges, the UH Graduate School, and UH Libraries devised a submission and approval process to implement this policy. Decentralized in nature, the UH ETD Program was designed to be distributed across colleges, the graduate school, and the libraries. Primary roles and responsibilities for colleges include making policies regarding content, structure, and deadlines; providing instruction and consultation to students on policies and document structure; and approving submitted documents based on localized policies and guidelines. The UH Graduate School enforces system-wide policies, including embargo requests and submission deadlines; compiles current lists of active departments and programs; and distributes submissions to appropriate colleges as part of the approval workflow. UH Libraries maintains the ETD submission software (Vireo); trains personnel at colleges and students to use the software; and releases documents to the institutional repository once embargoes expire. This shared approach has allowed stakeholders to accumulate over 3,200 ETDs to date.

The Libraries collaborate with the Texas Digital Library (TDL), a consortium of Texas higher education institutions focused on providing digital collections infrastructure, to administer two platforms to facilitate the ETD workflow process: Vireo and DSpace. Developed in 2009 by TDL and Texas A&M University with funding from the Institute of Museum and Library Services, Vireo is an open source software dedicated to managing the submission, approval, and publication of ETDs. The software provides an online submission module that collects user-supplied metadata and PDF versions of a student's thesis or dissertation. Upon submission, the Vireo platform tracks documents throughout the approval process, including verifications from the student's committee chair, from the college, and from the Libraries. After documents are fully vetted through all appropriate groups, they are released to UH's DSpace institutional repository. The metadata includes elements from both the Dublin Core Metadata Initiative terms namespace and custom elements outlined in the TDL guidelines, version 2. DSpace leverages the embedded optical character recognition text to make ETDs full-text searchable and freely available for search, download, and reuse.

Literature Review

The professional literature has been engaged with issues of metadata quality, metadata assessment, and the specific challenges of ETD metadata management for over two decades. The authors highlight some of the intersecting topics that informed their metadata remediation project and situate this case study in the larger practice of long-term metadata management.

Metadata quality has been explored by a number of researchers. In their influential paper, Bruce and Hillman acknowledge that what makes "good metadata" is often difficult to articulate and depends on its context.[2] They outline seven dimensions of metadata quality that can be applied generally to all metadata: completeness (chosen element set describes resources completely and elements are populated as fully as possible); accuracy (values are both factual and free of typographical errors); provenance (availability of contextual information about metadata creation and modification); conformance to expectations (elements and values fulfill target users' needs); logical consistency and coherence (standard element definition and input within and across collections); timeliness (metadata is up-to-date); and accessibility (open and available technologically and intellectually).[3] Tani, Candela, and Castelli surveyed the research on metadata quality frameworks and assessment techniques.[4] They summarize that "defining 'what metadata quality is'" is a very challenging task. It can be affirmed that no consensus has been reached on this concept until now, apart from the shared understanding that the difficulties in defining it come from its intrinsic characteristic of being a multidimensional and context-specific concept."[5]

Literature from the information profession also specifically addresses the management of ETD metadata quality. These conversations frequently address the challenges and opportunities that accompany the metadata creation process. In their case study of the metadata remediation process for the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), the University of Illinois at Urbana-Champaign's institutional repository, which includes ETDs, Stein, Applegate, and Robbins note that "Despite the existence of the Metadata Policy and Best Practices documentation, a variety of errors have been introduced into the IDEALS repository metadata via the

user-submission workflow and batch ingests of materials."[6] Researchers have identified metadata fields that are more likely to include errors and present long-term challenges for ETD management. Waugh et al., Lubas, and Chapman et al. have all addressed the challenges of managing controlled vocabularies in ETD collections.[7] Waugh et al. discuss the frequency with which metadata creators use various ways to express names in the repository, with each variation being ingested into the repository. These variations have implications for the discoverability of ETDs, as a user must know to search or browse for all instances of a name to obtain the desired documents. Waugh et al. also note that names play an important role in other ETD administrative functions, such as citation analysis and copyright and licensing management.[8] Chapman et al. state that the problem with names is compounded by the limited number of solutions available to institutions. They state that the

> Use of the Library of Congress Name Authority File is problematic because many authors in institutional repositories have no entry, as they tend to be authors of journal articles and conference papers, not books or monographs. Use of the campus-level directory can aid in some cases, but often faculty leave or publish under a name different from their directory name leaving gaps in its usefulness for authority control. There exists no standard to uniquely identify authors.[9]

Despite the identified limitations, Lubas discusses how the consistency of user-generated names improves when depositors are given controlled lists from which to choose.[10] Finally, Potvin and Thompson outline the challenges of managing a growing set of date metadata elements for ETDs.[11] They write that differing "philosophies about the role of metadata, viewed either as primarily descriptive or as a distinct component in the lifecycle management of electronic documents" have informed how dates are captured and expressed in metadata records.[12] Competing philosophies, in conjunction with repository software development, have caused date metadata to differ widely from prescribed ETD metadata standards (including the Networked Digital Library of Theses and Dissertations ETD-MS v1.1: An Interoperability Metadata Standard for Electronic Theses and Dissertations).[13]

Metadata quality is important, but its context-dependent nature makes it costly to assess. Some researchers have experimented with methods for automating metadata assessment. Nichols et al. compared two automated institutional repository metadata analysis tools: the Metadata Analysis Tool (MAT) from the University of Waikato and the Kiwi Research Information Service (KRIS) from the National Library of New Zealand. Both tools harvest metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and help metadata librarians analyze this data, pinpointing specific metadata errors and generating summary statistics.[14] Goovaerts and Leinders conducted a study on a random sample of OAI-PMH MODS metadata from the OceanDocs aggregated ocean research repository to statistically evaluate metadata quality.[15] In both cases, the statistical analyses were a useful tool for identifying errors and areas for improvement; however, context and thoughtful interpretation of statistical assessment results is required. Radio underscores the importance of closely analyzing statistical data used for metadata auditing purposes. Illustrating this, he notes the phenomenon of "data absence," which acknowledges that a metadata field devoid of a value is not, by default, inaccurate or incomplete.[16] Further complicating the metadata auditing process, Radio notes that "data absence" is just one "critical factor" that impacts "the interpretation of a metadata statement" during metadata auditing.[17] Consequently, automated assessment is best used when augmented by human intervention. Depending on the scale of the repository, manual assessment processes may be feasible. For example, Westbrook et al. used a random sampling method to audit metadata in the UH Digital Library according to Bruce and Hillman's quality framework summarized above.[18]

Statistical and other metadata analyses provide insights into data quality, which may inform metadata remediation efforts. At the UH Libraries, results from their metadata quality audit informed manual and automated remediation efforts to align digital collection metadata across collections.[19] As part of an effort to migrate to a new digital asset management system, Neatrour et al. performed limited metadata assessment and remediation and plan to pursue additional assessment and enhancement after the migration is complete.[20] Improving metadata is a time-consuming process that has implications for staffing resources and expertise. Moulaison Sandy and Dykas stress that the improvement of metadata quality can be increased by "adequate and appropriate staffing of the repository."[21] In other cases, it is not possible or desirable to dedicate resources to this work. Chapman, Reynolds, and Shreeves discuss the decision to forgo metadata remediation for the University of Illinois at Urbana-Champaign's (UIUC) institutional repository at that time because "it was not clear what the staffing implications were likely to be for the cataloging unit and due to chronic staffing shortages" and "there was a general feeling that because of the nature of the institutional repository, access to resources would principally occur through search engines and full text indexing."[22] Additionally, they note that a poor repository user interface, which fails to take advantage of batch processes, creates an extra burden on staff. Still, there are clear benefits of expanding resources for ETD metadata creation and remediation. McCutcheon

argues for the need to enhance ETD bibliographic records through mediation tasks, including "making sure that special characters are represented properly, doing name authority work, and subject analysis."[23] According to the author, this work will optimize the discoverability of ETDs, making them more widely available to those using library catalogs.[24] Since Chapman, Reynolds, and Shreeves' 2009 publication, UIUC has expanded staffing in Metadata Services, enabling them to undertake a metadata remediation project.[25]

Despite the growing literature on the benefits and limitations of metadata remediation, there are few case studies detailing the experiences of metadata review. In their paper on ETD metadata and quality control, Steele and Sump-Crethar note that "The issue of quality control is a topic worthy of an entire study. Our survey only asked about the importance and whether quality control was done."[26] They suggest that "Future research could examine further how quality control is done."[27] Focused on ETD metadata analysis and remediation, the authors' paper contributes one such case study, furthering the profession's understanding of metadata quality control processes.

Method

The Libraries initiated the ETD remediation process largely due to the release of TDL's revised ETD metadata standard.[28] The standard, initially developed in 2008 to assist with the aggregation of TDL members' ETDs through a statewide repository, articulates required and optional metadata elements needed to describe ETDs and make them accessible and discoverable via the web. While the first version of the standard addressed a wide array of metadata issues, the shifting nature of ETD submission software, coupled with emerging metadata areas popularized since the creation of the 2008 standard, including increased attention on author name disambiguation and explicit rights statements, prompted a revision to the standard.[29]

The revised standard included several changes that prompted UH Libraries to modify current practices and workflows. While the 2008 standard "centered around the Metadata Object Descriptive Standard (MODS) application profiles, with guidelines including flat, key-value paired Dublin Core (DC) and a thesis schema (known collectively as 'TDL DC') only for crosswalking to meet the Networked Digital Library of Theses and Dissertations (NDLTD) ETD-MS exchange standard," the 2015 revised standard is based on qualified DC, which more closely aligns with TDL members' current practices.[30]

The transition from MODS to a Qualified DC application profile required changes to certain metadata elements. For example, DC terminology like "Date" and "Format" replaced the MODS-related terms "Origin Information" and "Physical Description," respectively; values in some metadata fields were also better suited to other fields, including the transfer of URLs from "Location (URL)" to the "Identifier" element; and the removal of redundant fields, including values in "Record Information" since this information is automatically generated by DSpace and placed in administrative metadata fields (such as <dc.date.accessioned> and <dc.description.provenance>).[31] Beyond the shift from MODS to Qualified DC, additional changes promoted new and emerging aspects of ETD administration, such as rights metadata, author identifiers (ORCID), and description information, plus encoding guidelines to improve the discoverability of metadata in aggregated search platforms (e.g., Google Scholar's Highwire Press tags). Not all of the recommendation set out in TDL guidelines, version 2, were implemented by the project team; the following sections detail the specific issues that the authors addressed.

A team consisting of members of the Metadata Unit and the Head of Digital Research Services was formed in July 2015 to initiate the ETD metadata remediation project. Their charge was to develop a strategy focusing on reviewing the current state of the UH IR metadata, noting any deficiencies, and implementing a workflow to address any problems discovered while incorporating the latest best practices and adhering the recently developed TDL guidelines. The following sections detail the discrepancies the authors identified, the strategies and tools used to correct them, and the procedures followed.

Discrepancies

After exporting metadata from DSpace, an informal analysis of the CSV data in Microsoft Excel revealed the following issues, providing the foundation for the remediation (see table 1).

Strategies and Tools

Following the previous success with the UH Libraries metadata upgrade project, the authors adopted similar approaches for communication, documentation, and remediation to conduct the ETD metadata upgrade project.[32] The section below provides an overview of the strategies and tools used.

Communication

Communication is an integral part of the process. Since this was a complex project spanning a significant amount of time, the authors needed a means to communicate and collaborate internally. Basecamp, a project management

Table 1. Issues Found and Remediation Strategy

Issue	Notes	Example	Remediation Strategy
Duplicate Metadata Fields	Various metadata fields have similar information spread across duplicate columns in the exported CSV file.	dc.contributor.advisor dc.contributor.advisor[] dc.contributor.advisor[*]	Verify that the appropriate information was captured in the column with the correct field label (as outlined in the TDL guidelines). Metadata values need to be moved to one column so the duplicate columns could be removed.
Incorrect URLs in dc.identifier.uri field	Many records contained a faulty hyperlink that cluttered the user interface/display and confused our users.	Incorrect URL: http://hdl.handle.net/10657/ETD-UH-2010-05-34 Corrected URL: http://hdl.handle.net/10657/423	Review the entries in the dc.identifier.uri field and remove the incorrect url entries.
Inconsistent spellings for advisor and committee member names	The previous submission process allowed students to fill in free text fields with little or no moderation resulting in inconsistencies in spellings for advisor and committee member names.	Standardized form of name: Chou, Diana, S. Non-standardized form of name: Chou, Diana	Review the names and make sure each person has only one preferred form for their name.
Varying department and degree discipline names	To attract students to their constantly evolving fields of academic study, department, and discipline names are reevaluated and changed to reflect the latest trends and best practices. This resulted in inconsistencies in department and degree names across the ETD collection.	Legacy name: Educational Leadership and Cultural Studies Current name: Educational Leadership and Policy Studies	Confirm the correct form of names by conducting research and contacting college and department representatives.
Extra word "abstract" in the dc.description.abstract field for some ETD records	In this field we noticed the words "Abstract Abstract" and other formatting issues. This was likely a result of users cutting and pasting large amounts of text from their thesis into the submission form.	Correct Abstract: "In this study . . ." Incorrect Abstract: "Abstract. In this study . . ."	Delete the duplicate word "abstract."
Dates in various formats	Multiple date fields existed in our item records with many containing dates in various formats.	dc.date.created 2008-08 dc.date.issued 3/24/2010 dc.date.submitted 08-Aug	Update to current TDL standard.

platform, was selected to assign tasks, document and track decisions, and record meeting minutes. Additionally, communication with college and department stakeholders was necessary to complete their goal. Communicating regularly with a point of contact that has the institutional knowledge to answer questions about historical department and degree names allowed the team to address questionable data more effectively. The authors kept close contact with colleges and departments through email to communicate information externally.

Documentation

It was imperative to capture the remediation process to enable new team members to replicate the workflow necessary to continue this project. PMwiki, an open source wiki publishing platform, was used to document workflow processes, collect responses from personnel in colleges and departments, and archive project information. Team members frequently used both screenshots and step-by-step descriptions to ensure that the instructions are easy to understand and usable for future reference.

Remediation

Remediation entailed making the necessary metadata edits and corrections to align content in the IR with the newly updated TDL guidelines, version 2.[33] Microsoft Access and OpenRefine were chosen since the authors were familiar with these tools and could use them to automate portions of the workflow, reducing repetitive tasks and human error. Microsoft Access queries are useful to perform complex

functions such as consolidating values from multiple columns or cells. OpenRefine was a great asset to standardize author, advisor, department, and college names with the facet, filter, and cluster functions.

ETD Metadata Workflow

Based on the issues identified during the export analysis (outlined in table 1), the authors initiated the remediation workflow. Metadata for the then 900+ ETDs was exported from the DSpace repository as a CSV file and opened in OpenOffice, which retains any special character encoding found in the metadata.[34] It was then saved as a Microsoft Excel .xls file and imported into Access for remediation work.

Remove Duplicate Columns

In the exported CSV file, duplicate columns were found that represented a single metadata element. For example, three columns contained values for the thesis advisor: dccontributoradvisor, dccontributoradvisor1, and dccontributoradvisor2 (see figure 1).[35]

Before the authors could perform remediation work, they first consolidated the values across duplicate columns into a single column. An update query (see figure 2) was used to copy the values from one column to another (see figure 3).

A simple "copy and paste" command should accomplish the task; however, using the update query minimizes human errors. After the values were in one column, that data was ready to be edited.

Remove Incorrect URLs from dc.identifier.uri Field

The authors also identified broken URLs in the dc.identifier.uri field (see figure 4). Because the correct URLs are all of the same character length, they were able to use Access's "right" function to retain the correct URLs in the column while removing the incorrect URLs. This function allowed them to retain the *x* number of characters from the right, in this case the thirty-one characters (which is the length of the correct urls) from the right.

Figure 5 shows the update query to complete the task. Figure 6 shows the query result.

Name Standardization

The authors identified inconsistent forms of names throughout the dc.contributor.advisor and dc.contributor.committeemember columns. To ensure one preferred form of name for each person, the authors imported the columns that contain advisor and committee member names with the record ID and collection ID into OpenRefine for name standardization (see figure 7).

DSpace Export 09-01-15

id	collection	dccontributoradvisor	dccontributoradvisor1	dccontributoradvisor2
3	10657/2			
10	10657/2			
22	10657/2			Phillips, Scott
24	10657/2	Lee, Rebecca E.		
157	10657/2		Arbona, Consuelo	
159	10657/2		Arbona, Consuelo	
160	10657/2		Coleman, Nicole	
161	10657/2		Chow, Diana	
162	10657/2		Ryan, Michael	
165	10657/2		Yamasaki, Jill	
167	10657/2		Schwartz, Jonathan	
168	10657/2		Mountain, Lee	
169	10657/2		Armsworth, Mary	
170	10657/2		Arbona, Consuelo	
171	10657/2		Olson, Beth	
172	10657/2		Jowett, Garth	
173	10657/2		Ryan, Michael	
174	10657/2		Vardeman-Winter, Jennifer	
175	10657/2		Busch, Steven D.	
176	10657/2		MacNeil, Angus J.	
177	10657/2		Craig, Cheryl J.	
178	10657/2		White, Cameron	
179	10657/2		Vardeman-Winter, Jennifer	

Figure 1. Exported xls file in Microsoft Access with duplicate columns

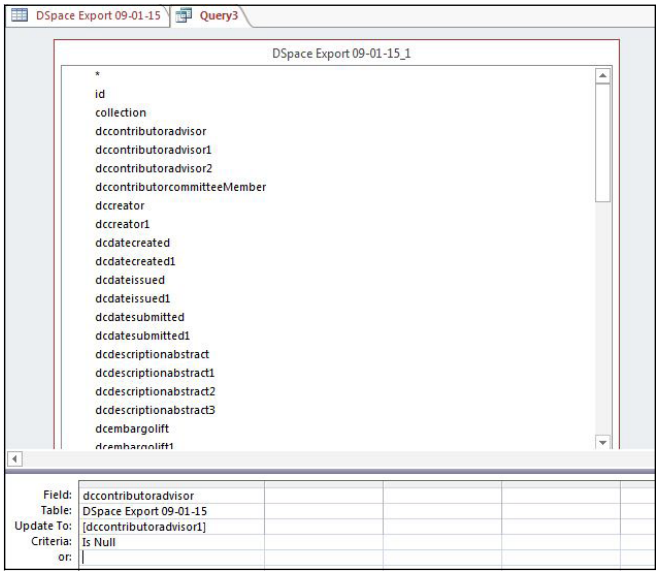
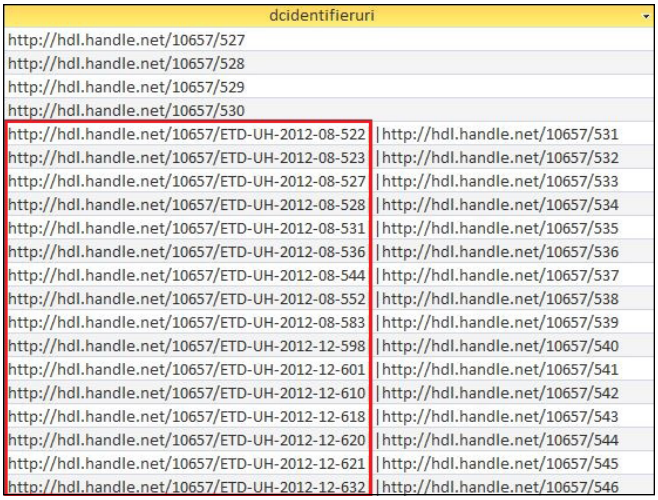


Figure 2. Update query in Access

DSpace Export 09-01-15

id	collection	dccontributoradvisor	dccontributoradvisor1	dccontributoradvisor2
3	10657/2			
10	10657/2			
22	10657/2			
24	10657/2	Lee, Rebecca E.		
157	10657/2		Arbona, Consuelo	
159	10657/2		Arbona, Consuelo	
160	10657/2		Coleman, Nicole	
161	10657/2		Chow, Diana	
162	10657/2		Ryan, Michael	
165	10657/2		Yamasaki, Jill	
167	10657/2		Schwartz, Jonathan	
168	10657/2		Mountain, Lee	
169	10657/2		Armsworth, Mary	
170	10657/2		Arbona, Consuelo	
171	10657/2		Olson, Beth	
172	10657/2		Jowett, Garth	
173	10657/2		Ryan, Michael	
174	10657/2		Vardeman-Winter, Jennifer	
175	10657/2		Busch, Steven D.	
176	10657/2		MacNeil, Angus J.	
177	10657/2		Craig, Cheryl J.	
178	10657/2		White, Cameron	
179	10657/2		Vardeman-Winter, Jennifer	

Figure 3. Results of update query shown in figure 2



dcidentifieruri	
http://hdl.handle.net/10657/527	
http://hdl.handle.net/10657/528	
http://hdl.handle.net/10657/529	
http://hdl.handle.net/10657/530	
http://hdl.handle.net/10657/ETD-UH-2012-08-522	http://hdl.handle.net/10657/531
http://hdl.handle.net/10657/ETD-UH-2012-08-523	http://hdl.handle.net/10657/532
http://hdl.handle.net/10657/ETD-UH-2012-08-527	http://hdl.handle.net/10657/533
http://hdl.handle.net/10657/ETD-UH-2012-08-528	http://hdl.handle.net/10657/534
http://hdl.handle.net/10657/ETD-UH-2012-08-531	http://hdl.handle.net/10657/535
http://hdl.handle.net/10657/ETD-UH-2012-08-536	http://hdl.handle.net/10657/536
http://hdl.handle.net/10657/ETD-UH-2012-08-544	http://hdl.handle.net/10657/537
http://hdl.handle.net/10657/ETD-UH-2012-08-552	http://hdl.handle.net/10657/538
http://hdl.handle.net/10657/ETD-UH-2012-08-583	http://hdl.handle.net/10657/539
http://hdl.handle.net/10657/ETD-UH-2012-12-598	http://hdl.handle.net/10657/540
http://hdl.handle.net/10657/ETD-UH-2012-12-601	http://hdl.handle.net/10657/541
http://hdl.handle.net/10657/ETD-UH-2012-12-610	http://hdl.handle.net/10657/542
http://hdl.handle.net/10657/ETD-UH-2012-12-618	http://hdl.handle.net/10657/543
http://hdl.handle.net/10657/ETD-UH-2012-12-620	http://hdl.handle.net/10657/544
http://hdl.handle.net/10657/ETD-UH-2012-12-621	http://hdl.handle.net/10657/545
http://hdl.handle.net/10657/ETD-UH-2012-12-632	http://hdl.handle.net/10657/546

Figure 4. Broken URLs in the dc.identifier.uri field

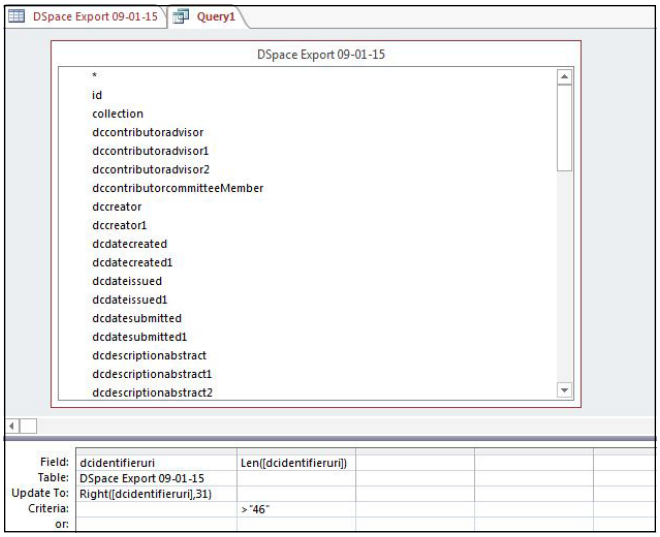
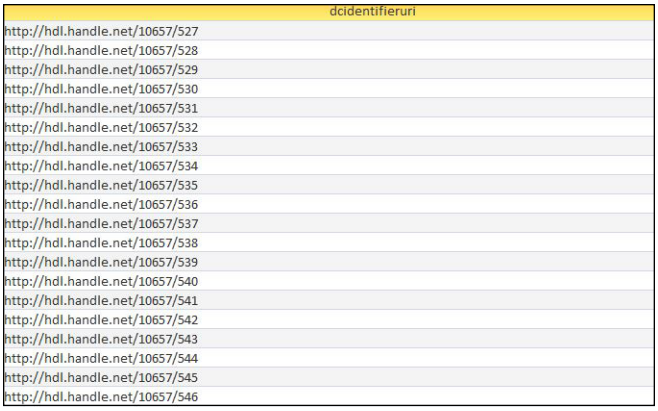


Figure 5. Update query to remove broken URLs



dcidentifieruri
http://hdl.handle.net/10657/527
http://hdl.handle.net/10657/528
http://hdl.handle.net/10657/529
http://hdl.handle.net/10657/530
http://hdl.handle.net/10657/531
http://hdl.handle.net/10657/532
http://hdl.handle.net/10657/533
http://hdl.handle.net/10657/534
http://hdl.handle.net/10657/535
http://hdl.handle.net/10657/536
http://hdl.handle.net/10657/537
http://hdl.handle.net/10657/538
http://hdl.handle.net/10657/539
http://hdl.handle.net/10657/540
http://hdl.handle.net/10657/541
http://hdl.handle.net/10657/542
http://hdl.handle.net/10657/543
http://hdl.handle.net/10657/544
http://hdl.handle.net/10657/545
http://hdl.handle.net/10657/546

Figure 6. Query results from figure 5

All	id	collection	dccontributoradvisor	dccontributorcommitteeMember
1.	3	10657/2	Bond, Richard A.	Bond, Richard A.||Knoll, Brian J.||Pedemonte, Carlos H.||Cahill, Gregory M.||Moore, Robert H.
2.	10	10657/2	Chow, Diana	Chow, Diana||Bates, Theodore R.||Hu, Ming||Liang, Dong||Tam, Vincent H.
3.	22	10657/2	Phillips, Scott	
4.	24	10657/2	Lee, Rebecca E.	Layne, Charles S.||O'Connor, Daniel P.||Rifai, Hanadi
5.	157	10657/2	Arbona, Consuelo	Burridge, Andrea||Gaa, John||Backus, Margot
6.	159	10657/2	Arbona, Consuelo	Day, Susan X.||Armsworth, Mary||Olvera, Norma
7.	160	10657/2	Coleman, Nicole M.	Arbona, Consuelo||Mc Pherson, Robert||Watson, Margaret
8.	161	10657/2	Chow, Diana	Yeung, Jim||Bond, Richard||Liang, Dong||Giovanella, Bappino
9.	162	10657/2	Ryan, Michael	Vardeman, Jennifer||Lewis, Dontee
10.	165	10657/2	Yamasaki, Jill	Haun, Martha||Addison Otey, Penny
11.	167	10657/2	Schwartz, Jonathan	Arbona, Consuelo||Wiesner, Margit||Olvera, Norma
12.	168	10657/2	Mountain, Lee	Abrahamson, Richard||Craig, Cheryl||Horn, Catherine
13.	169	10657/2	Armsworth, Mary	Greer, Gary||Gaa, John||Dao, Tam||Andrews, Nicole
14.	170	10657/2	Arbona, Consuelo	Watson, Margaret||Coleman, Nicole||Armsworth, Mary
15.	171	10657/2	Olson, Beth	Haun, Martha||Verheyen, Claremarie
16.	172	10657/2	Jowett, Garth	Houk, Keith R.||Reed, John

Figure 7. Imported names in OpenRefine

Figure 8. "Split multi-valued cells" command

They next divided committee member names into separate rows using the "split multi-valued cells" command (see figure 8).

The authors consolidated all advisor and committee member names into one column to standardize names from a single list by using the "transpose cells across columns into rows" command (see figures 9 and 10).

This function also enabled the authors to track the field from which a value originated and store this location in an additional column, "Original Column." Tracking allowed them to return the standardized names to their original fields after name cleaning was complete. Figure 11 shows the result of using the "transpose cells across columns into rows" command: all advisor and committee member names are in one column, allowing the authors to use "facet and cluster" commands to standardize the names (see figures 12 and 13).

After name standardization was locally completed, the authors reconciled this list with the Library of Congress Name Authority File (LCNAF) and updated any existing names in OpenRefine to reflect LCNAF values.

Following reconciliation, the authors separated the advisor and committee member names back into two columns and imported them into Access. They first used the "text filter" command in OpenRefine to display only advisor names (see figure 14).

Next they used the "Add column based on this column" command to create a new column, dccontributoradvisor, for the advisor values (see figures 15 and 16). The authors

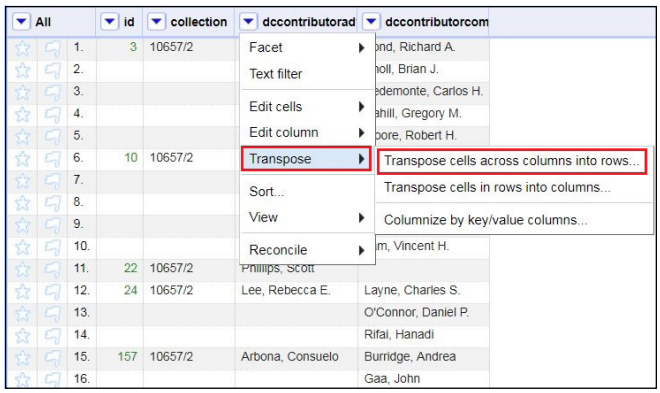


Figure 9. "Transpose cells across columns into rows" command

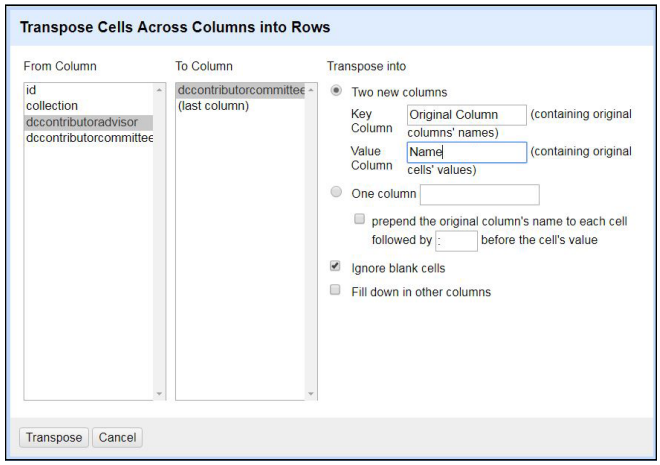


Figure 10. Adjust settings to "transpose cells across columns into rows"

applied the same steps to pull committee member names into a new column. The text filter was removed to view the results (see figure 17). After there were two new columns for dccontributoradvisor and dccontributorcommitteemember, the authors deleted the previous two columns created just for name standardization. To import the standardized name back into Access, they placed multiple name values into a single cell using the "join multi-valued cells" command (see figures 18 and 19). The authors deleted any remaining empty rows by filtering for blank cells and removing them from the table. The final table was then ready to be imported back to Access (see figure 20).

Using OpenRefine, the authors exported the file as an Excel spreadsheet and imported it into Access as a new table. Since each record has a unique ID, they used Access's "join table" function to combine the new table with the existing one, shown in figure 21. The authors then deleted the columns containing the original advisor and committee member names, concluding the name cleanup process.

All	id	collection	Original Column	Name
1.	3	10657/2	dccontributoradvisor	Bond, Richard A.
2.			dccontributorcommitteeMember	Bond, Richard A.
3.			dccontributorcommitteeMember	Knoll, Brian J.
4.			dccontributorcommitteeMember	Pedemonte, Carlos H.
5.			dccontributorcommitteeMember	Cahill, Gregory M.
6.			dccontributorcommitteeMember	Moore, Robert H.
7.	10	10657/2	dccontributoradvisor	Chow, Diana
8.			dccontributorcommitteeMember	Chow, Diana
9.			dccontributorcommitteeMember	Bates, Theodore R.
10.			dccontributorcommitteeMember	Hu, Ming
11.			dccontributorcommitteeMember	Liang, Dong
12.			dccontributorcommitteeMember	Tam, Vincent H.
13.	22	10657/2	dccontributoradvisor	Phillips, Scott
14.	24	10657/2	dccontributoradvisor	Lee, Rebecca E.
15.			dccontributorcommitteeMember	Layne, Charles S.
16.			dccontributorcommitteeMember	O'Connor, Daniel P.
17.			dccontributorcommitteeMember	Rifai, Hanadi
18.	157	10657/2	dccontributoradvisor	Arbona, Consuelo
19.			dccontributorcommitteeMember	Burridge, Andrea
20.			dccontributorcommitteeMember	Gaa, John

Figure 11. Result of the "transpose cells across columns into rows" command

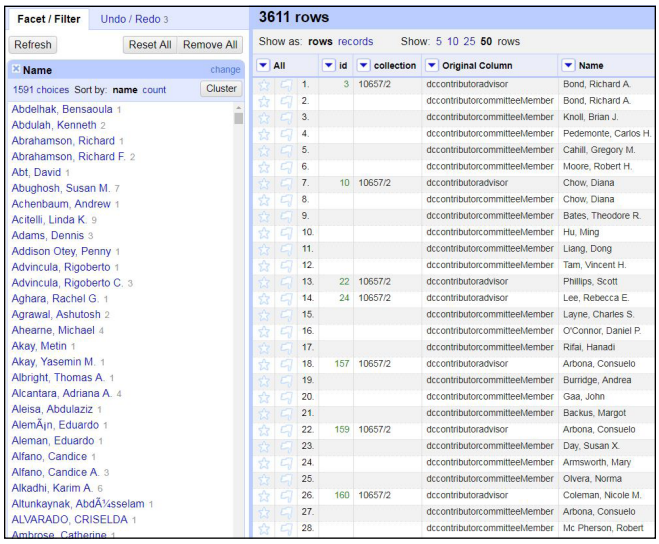


Figure 12. "Facet" command in OpenRefine

Additional Standardization Tasks

The authors filtered for all records in which the value in the Abstract field began with "Abstract" (see figure 22) and manually deleted this word. Access's "sort" function was used to sort the department and discipline names and confirmed the accuracy of these names with respective colleges and departments.

In compliance with the new TDL guidelines, version 2, the authors deleted the original dcdateissued column that contained the dates in YYYY-MM-DD format. Using

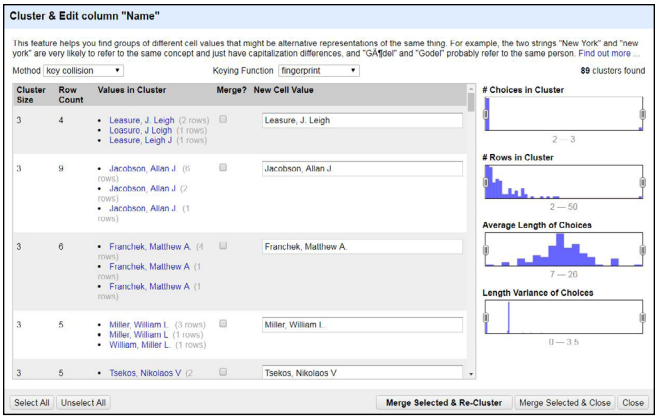


Figure 13. "Cluster & Edit" command in OpenRefine

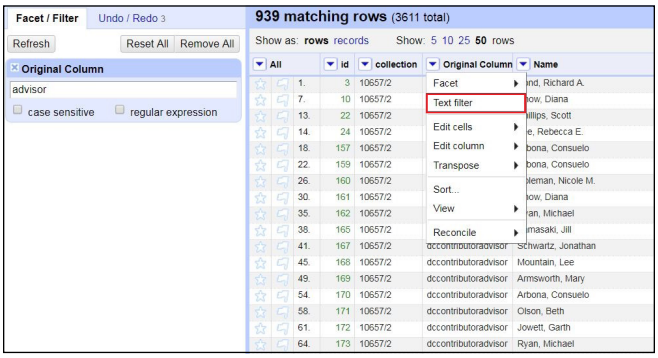


Figure 14. "Text filter" command

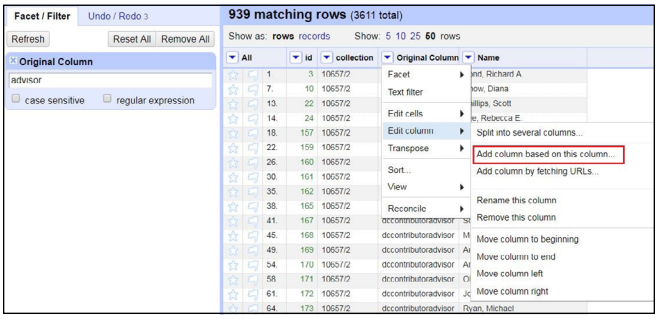


Figure 15. "Add column based on this column" command

Access, they renamed and reformatted values in two other date columns: dcdateissued (YYYY-MM) and dcdatecreated (Month Year) (see figure 23).

When these tasks were complete, the authors exported the updated Access file to Excel, imported the file into OpenOffice (to retain special character encoding), and saved it as a CSV file. Finally, the CSV file was ingested into DSpace.

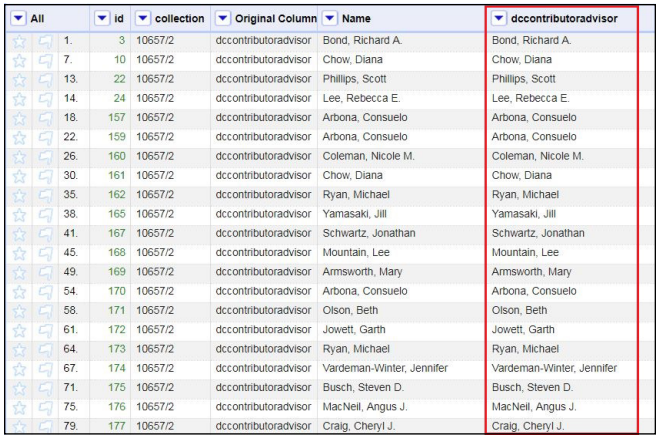


Figure 16. Result of "Add column based on this column" command

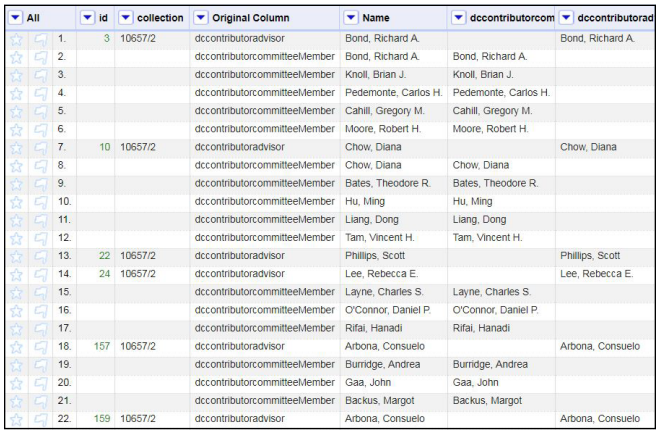


Figure 17. Final result showing two newly created columns for standardized advisor and committee member names

Discussion

Throughout the year-long project, the authors encountered situations that required them to make local decisions about editing specific metadata fields. They investigated how to integrate external tools to reduce future errors in metadata creation and maintenance. In the following section, they outline lessons learned, the project's limitations, and the project's next steps.

Lessons Learned

While undertaking the ETD upgrade project, the authors determined which required and optional metadata fields from the TDL guidelines, version 2, to implement.[36] They elected to omit optional fields, such as dc.embargo dc.format.extent and dc.subject.lcsh.[37] Because of the complexities surrounding the optional dc.rights field, the authors elected to upgrade this field at a later date.[38] They

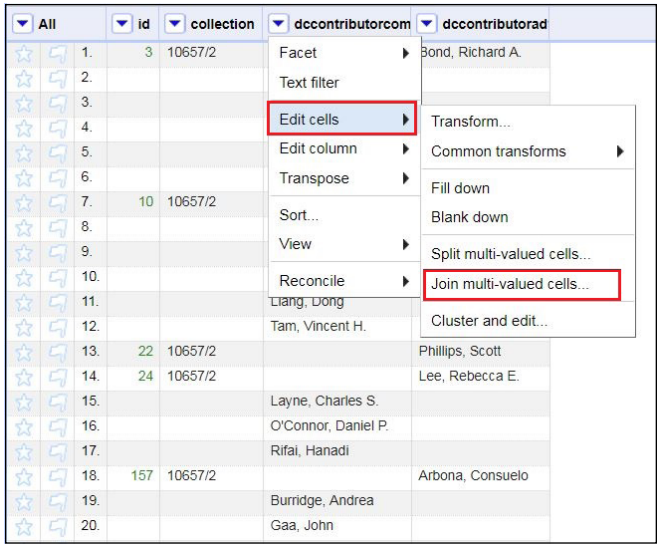


Figure 18. "Join multi-valued cells" command

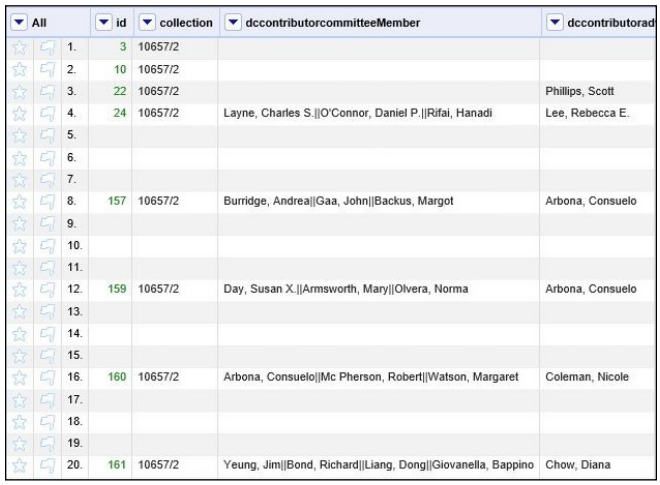


Figure 19. Result after deletion of original two columns for name standardization and use of "join multi-valued cells" command for dccontributorcommiteemember column

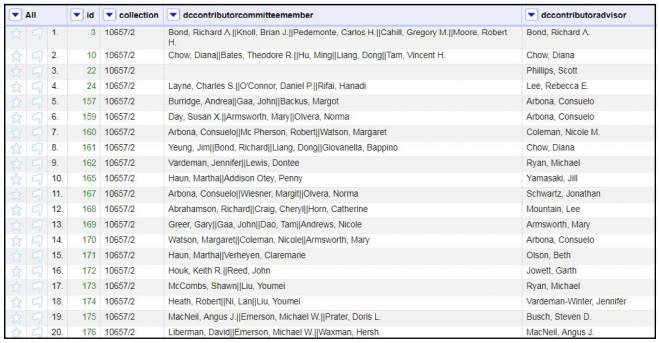


Figure 20. Final results

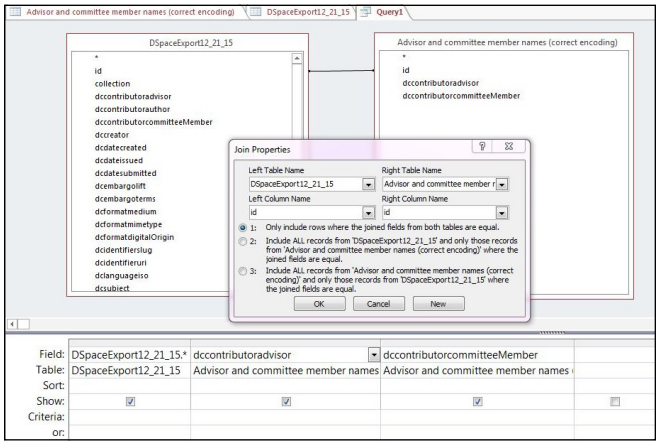


Figure 21. Configure settings—"join properties for join tables" function in Access

also placed emphasis on retaining fields that added value to staff or users. For example, dc.date.accessioned, which can be used by staff to determine whether records had been remediated in a previous batch, was retained. Dc.date.accessioned records the date the DSpace repository first receives the thesis; this value does not change after remediating and reloading the metadata.

The team also made decisions about the level of quality for metadata records. Adhering to Voltaire's maxim that "the best is the enemy of the good," they passed on opportunities to make perfect records to complete the project.[39] For example, some students submitted ETDs with values in all capital letters. These records were not changed because this formatting does not affect searching and standardizing case is not a priority.

During the re-ingest process, the authors discovered that DSpace required them to retain the same number of elements and element labels originally exported; otherwise DSpace would not recognize the edits.[40] They were limited to ingesting one hundred records at a time due to TDL's system configuration. The authors discovered that including the dc.description.abstract field in the re-ingest process caused errors, and manually edited this field after re-ingesting content.

To capitalize on the process of name standardization for people, departments, and degree disciplines, the authors compiled a set of local controlled vocabularies: advisor and committee member names and department and degree names. Controlled vocabularies would reduce both the user-generated errors that occur when students are inputting information in the free text fields and the staff time needed to remediate future batches. The authors used the reconcile-CSV software to implement the local-controlled personal name vocabulary.[41] This tool is used in

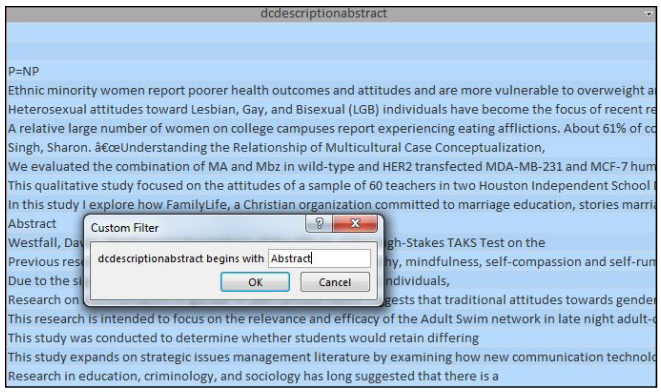


Figure 22. Using the "filter" function in Access to remove extra word "abstract" from the abstract field

dcdateissued	dcdatecreated
2008-08	August2018
2008-08	August2018
	May2018
2010-05	May2018
2010-08	August2018
2011-05	May2018
2010-08	August2018
2010-12	December2018
2010-12	December2018
2010-12	December2018
2010-08	August2018
2010-12	December2018
2010-12	December2018
2010-12	December2018
2010-12	December2018
2010-12	December2018

Figure 23. Result showing dcdateissued and dcdatecreated columns in Access

conjunction with OpenRefine and allows users to reconcile project data against data from a CSV file to authorize and standardize values. For the local department and discipline name vocabularies, the authors supplied a dropdown list of verified values in Vireo. Continual maintenance of their local-controlled vocabularies requires minimal resources, with the greatest demand from the capture of new values in each subsequent batch.

Another key lesson was the critical importance of communication between the metadata remediation team, the graduate school, and other university colleges and departments. The frequent changes of department names and degree discipline names created confusion and inconvenience during the remediation. An established communication channel with counterparts from various colleges and departments provided firsthand information and enabled the authors to track down past changes and save efforts for future cleanup work.

Limitations

This case study provides metadata practitioners with one strategy for remediation, but it is important to consider the type, scale, and peculiarities of a particular project before employing remediation strategies. The transformations the authors performed using Access and Open Refine worked well for the scale of their project. However, these processes may not be appropriate for institutions working on a larger scale project, for example ten thousand records or more. Other approaches, such as scripting, may work better in these instances. Stein, Applegate, and Robbins note their use of scripts for metadata remediation of works in the IDEALS repository.[42] This technique was more appropriate for their strategy to "[remediate] values of a particular metadata field across multiple collections and communities when they do not match specified IDEALS best practices."[43] This differs from the authors' strategy to remediate all values from all works of the ETD community. The authors' strategy would also be appropriate for those undertaking remediation efforts who lack experience creating or using scripts. Additionally, while Access is not ideal for larger batches, OpenRefine supports the review and revision of larger CSV files of twelve thousand to ninety thousand rows.[44]

Another limitation is that the authors used an externally hosted repository. Although these tools are open source, thus providing flexibility in terms of customization and extensibility, they are hosted by TDL and not locally, and the authors lack direct access to the source code and data to implement scripts and other automation to enable further efficiencies.

Conclusion

In this case study, the authors developed sustainable workflows to bring their ETDs into compliance with an updated metadata standard. After completing the remediation process for all ETDs added to the IR between 2011 and 2015, the authors reviewed and finalized the documentation created during the process to replicate the process for future batches. They use this remediation workflow for each new batch of approximately two hundred to three hundred ETDs ingested into the repository two to three times a year.

While the scale is smaller compared to some institutions, and the authors are using an externally hosted platform, they plan to explore automated options for remediating future ETD deposits. These efforts include developing scripts to automatically manipulate values in the DSpace export file and name reconciliation in OpenRefine using their locally developed linked data vocabulary manager.[45]

The authors' case study joins a growing body of metadata remediation projects, including previous work discussed in the literature review. Examining these isolated case studies will begin to yield critical comparisons across projects, including the motivations for metadata remediation, the scope and methods used to conduct audits and data cleaning, and the resources and expertise needed to successfully complete such initiatives. This cross-sectional analysis would benefit a growing professional interest in and need for metadata remediation guidelines and common practices.

References and Notes

1. Texas Digital Library, "Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations" Version 2.0, September 2015, http://hdl.handle.net/2249.1/68437; Kara Long (Baylor University), Colleen Lyon (University of Texas), Kristi Park (Texas Digital Library), Monica Rivero (Rice University), and Santi Thompson (UH Libraries) were members of the 2015 working group. Sarah Potvin (Texas A&M University) chaired the working group.
2. Thomas R. Bruce and Diane I. Hillman, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice*, ed. Diane I. Hillman and Elaine L. Westbrooks (Chicago: American Library Association, 2004), 238–56.
3. Ibid., 243–49.
4. Alice Tani, Leonardo Candela, and Donatella Castelli, "Dealing with Metadata Quality: The Legacy of Digital Library Efforts," *Information Processing & Management* 49, no. 6 (2013): 1194–205.
5. Ibid., 1195.
6. Ayla Stein, Kelly J. Applegate, and Seth Robbins, "Achieving and Maintaining Metadata Quality: Toward a Sustainable Workflow for the IDEALS Institutional Repository," *Cataloging & Classification Quarterly* 55, nos. 7–8 (2017): 1–23, https://doi.org/10.1080/01639374.2017.1358786.
7. Laura Waugh, Hannah Tarver, and Mark Edward Phillips, "Introducing Name Authority into an ETD Collection," *Library Management* 35, nos. 4–5 (2014): 271–83; Rebecca L. Lubas, "Defining Best Practices in Electronic Thesis and Dissertation Metadata," *Journal of Library Metadata* 9, nos. 3–4 (2009): 252–63.; John W. Chapman, David Reynolds, and Sarah A. Shreeves, "Repository Metadata: Approaches and Challenges," *Cataloging & Classification Quarterly* 47, nos. 3–4 (2009): 309–25.
8. Waugh, Tarver, and Phillips, 272–73.
9. Chapman, Reynolds, and Shreeves, "Repository Metadata," 311.
10. Lubas, "Defining Best Practices in Electronic Thesis and Dissertation Metadata," 260.
11. Sarah Potvin and Santi Thompson, "An Analysis of Evolving Metadata Influences, Standards, and Practices in Electronic Rheses and Dissertations," *Library Resources & Technical Services* 60, no. 2 (2016): 99–114.
12. Ibid., 100.
13. Potvin and Thompson, "An Analysis of Evolving Metadata Influences," 107–8; Networked Digital Library of Theses and Dissertations, "Metadata, ETD-MS v1.1: An Interoperability Metadata Standard for Electronic Theses and Dissertations," ed. Thom Hickey, Ana Pavani, and Hussein Suleman, accessed January 31, 2018, www.ndltd.org/standards/metadata.
14. David M. Nichols et al., "Experiences in Deploying Metadata Analysis Tools for Institutional Repositories," *Cataloging & Classification Quarterly* 47, nos. 3–4 (2009): 229–48.
15. Marc Goovaerts and Dirk Leinders, "Metadata Quality Evaluation of a Repository Based on a Sample Technique," in *Research Conference on Metadata and Semantic Research*, ed. Juan Manuel Dodero, Manuel Palomo-Duarte, and Pythagoras Karampiperis (Berlin: Springer, 2012), 181–89.
16. Erik Radio, "Semiotic Principles for Metadata Auditing and Evaluation," *Cataloging & Classification Quarterly* 54, no. 2 (2016): 117–35.
17. Ibid., 25–26.
18. R. Niccole Westbrook et al., "Metadata Clean Sweep: A Digital Library Audit Project," *D-Lib Magazine* 18, nos. 5–6 (2012), https://doi.org/10.1045/may2012-westbrook; Bruce and Hillman, 243–49.
19. R. Niccole Westbrook et al., detail the metadata quality audit project. Discussion of the metadata remediation efforts included in Andrew Weidner, Annie Wu, and Santi Thompson, "Automated Enhancement of Controlled Vocabularies: Upgrading Legacy Metadata in CONTENTdm," in *International Conference on Dublin Core and Metadata Applications*, 2014, 167–72, http://dcpapers.dublincore.org/pubs/article/view/3716/1939; Santi Thompson and Annie Wu, "Metadata overhaul: upgrading metadata in the University of Houston Digital Library," *Journal of Digital Media*

Management 2, no. 2 (2013): 137–47.
20. Anna Neatrour et al., "A Clean Sweep: The Tools and Processes of a Successful Metadata Migration," *Journal of Web Librarianship* 11, nos. 3–4 (2017): 194–208, https://doi.org/10.1080/19322909.2017.1360167.
21. Heather Moulaison Sandy and Felicity Dykas, "High-quality Metadata and Repository Staffing: Perceptions of United States–Based OpenDOAR Participants," *Cataloging & Classification Quarterly* 54, no. 2 (2016): 114.
22. Chapman, Reynolds, and Shreeves, 320–22.
23. Sevim McCutcheon, "Basic, Fuller, Fullest: Treatment Options for Electronic Theses and Dissertations," *Library Collections, Acquisitions, & Technical Services* 35, nos. 2–3 (2011): 65.
24. Ibid., 65.
25. Stein, Applegate, and Robbins, "Achieving and Maintaining Metadata Quality," 8.
26. Tom Steele and Nicole Sump-Crethar, "Metadata for Electronic Theses and Dissertations: A Survey of Institutional Repositories," *Journal of Library Metadata* 16, no. 1 (2016): 66, https://doi.org/10.1080/19386389.2016.1161462.
27. Ibid., 66.
28. Texas Digital Library, "Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations" Version 2.0, accessed January 31, 2018, https://tdl-ir.tdl.org/handle/2249.1/68437.
29. Texas Digital Library, "Report for Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations," Version 2.0, September 2015, 3–4, http://hdl.handle.net/2249.1/68437.
30. Ibid., 4.
31. Texas Digital Library, "Dictionary of Texas Digital Library Descriptive Metadata," accessed January 31, 2018.
32. Weidner, Wu, and Thompson, 167-172.
33. Texas Digital Library, "Dictionary of Texas Digital Library Descriptive Metadata," accessed January 31, 2018.
34. The authors elected not to use Microsoft Excel because it does not consistently support UTF-8 special character/diacritic encoding in CSV files.
35. Note that fields have been renamed due to name conventions in Access. The original field names for dccontributoradvisor1 and dccontributoradvisor2 are dccontributoradvisor[] and dccontributoradvisor[*]. The reader should also note that column names in Access do not represent multiple values in a metadata field. Multiple values in a single element are represented in a list separated by the "||" symbol.
36. Texas Digital Library, "Dictionary of Texas Digital Library Descriptive Metadata," accessed January 31, 2018.
37. The dc.embargo field provides information on metadata or files that have temporary access restrictions. The authors chose not to implement this field because their local practice for embargoed items was to not ingest them into the repository, eliminating the need for this field.; dc.format.extent frequently describes the number of physical pages in an ETD. Since this information (a) is not supplied by the depositor; and (b) is not used to facilitate browsing or discoverability, the authors saw no value including it; dc.subject.lcsh provides additional descriptions of the content using terms from the Library of Congress Subject Headings (LCSH). As the ETD remediation process proceeded, they added LCSH to their ETD records when found in the pre-existing catalog record with the assistance of a cataloging librarian. In retrospect, this was a significant commitment in terms of time and labor. It required both the cataloging librarian to assign LCSH for each ETD record and, afterwards, the metadata specialist to update the remediation file with the newly assigned LCSH. This additional work forced the team to question whether adding LCSH would significantly help user searches. After careful deliberation, the authors decided not to repeat the effort for future batches since DSpace offers full-text indexing functionality that should enable users to apply keyword searches and find what they need.
38. The dc.rights field provides information on the rights "held in and over the resource" ("Dictionary of TDL Guidelines," p. 15). Determining the values used to populate this field can be labor intensive. While most ETDs are copyrighted by the student, some contain materials that students have previously published prior to submitting their ETD for graduation. In these cases, additional work is needed to document the correct citation/attribution for the previously published works. The authors are currently establishing a workflow to articulate the rights status of all ETDs in the collection.
39. Voltaire, *La bégueule. Conte moral* (Geneve, 1772).
40. For example, the authors kept all the corrected advisor names in the dc.contributor.advisor column and left dc.contributor.advisor[], dc.contributor.advisor[*] empty.
41. http://okfnlabs.org/reconcile-csv/.
42. Stein, Applegate, and Robbins, "Achieving and Maintaining Metadata Quality," 9.
43. Ibid., 10.
44. Ibid., 11.
45. For more information on this linked data vocabulary manager, see Andrew Weidner et al., "Outside the Box: Building a Digital Asset Management Ecosystem for Preservation and Access," *Code4Lib Journal* 36 (2017), accessed January 31, 2018, http://journal.code4lib.org/articles/12342.