Report for Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2

September 2015

Prepared by the TDL ETD Metadata Working Group
Sarah Potvin, Chair (Texas A&M University)
Kara Long (Baylor University)
Colleen Lyon (University of Texas)
Kristi Park (Texas Digital Library)
Monica Rivero (Rice University)
Santi Thompson (University of Houston)

Note: This report discusses the process and findings that informed the creation of the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2* and should be used in tandem with that document.

Report for Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2

September 2015

Table of Contents

Introduction and rationale

Vireo, the Texas Digital Library (TDL)-hosted Electronic Thesis and Dissertation submission and management application, has been enjoying a long moment of success. A growing number of TDL member institutions as well as institutions outside of TDL and Texas, including Harvard, the University of Illinois, and Johns Hopkins have adopted Vireo. When the tool was deployed as a prototype in 2007, it based metadata creation on the TDL ETD MODS application profile published in 2005.[1] Thanks to more than $800,000 in underwriting from a 2007-2010 Institute of Museum and Library Services (IMLS) National Leadership Grant, the TDL was able to develop that prototype into Vireo. Launched in 2008, Vireo relied on descriptive metadata guidelines produced by a TDL working group that year.[2]

Vireo has expanded and flourished, meeting the needs of diverse stakeholders in libraries and graduate offices, without close oversight from metadata stakeholders. Accordingly, the gap between the formal metadata guidelines that provided the basis for the tool's creation, and the metadata that the tool actually produces-- itself constituting a de facto standard-- has widened considerably.

Recognition of this widening gap prompted the formation, in 2014, of the Texas Digital Library Electronic Thesis and Dissertation Metadata Working Group, with members Kara Long (Baylor University), Colleen Lyon (University of Texas at Austin), Kristi Park (Texas Digital Library), Sarah Potvin (Texas A&M University), Monica Rivero (Rice University), and Santi Thompson (University of Houston). The TDL ETD working group was charged with providing "guidance to TDL member institutions and other ETD practitioners on metadata for electronic theses and dissertations, with a particular focus on works published through … Vireo." This guidance would take the form of an update to the 2008 TDL "Descriptive Metadata Guidelines for Electronic Theses and Dissertations"; an evaluation of how Vireo functionality diverged from metadata guidelines (and recommendations for correcting any divergence); and the publication and promotion of these updated guidelines.

The divergence of Vireo-generated metadata from TDL's metadata standard represented a lag in the larger system. A driving rationale for the development of Vireo had been its ability to support and enforce the creation of consistent metadata for ETDs, and to bring TDL member institutions-- and any other users of the application-- into shared practices and standards.

---

[1] Texas Digital Library, "MODS Application Profile for Electronic Theses and Dissertations," (Austin, TX: Texas Digital Library, 2005). Retrieved August 5, 2015 from http://www.tdl.org/wp-content/uploads/2009/04/etd_mods_profile.pdf The working group that produced the application profile included Brian E. Surratt-- Chair (Texas A&M University), Alisha Little (University of Texas), Anne M. Mitchell (University of Houston), and Jason Thomale (Texas Tech University).

[2] Texas Digital Library, "Descriptive Metadata Guidelines for Electronic Theses and Dissertations," Version 1.0 (June 2008). Retrieved August 15, 2015 from http://www.tdl.org/wp-content/uploads/2009/04/tdl-descriptive-metadata-guidelines-for-etd-v1.pdf Members of the working group: Amy Rushing--Chair (University of Texas), Jay Koenig (Texas A&M University), Anne Mitchell (University of Houston), William Moen (University of North Texas), Tim Strawn (University of Texas), and Jason Thomale (Texas Tech University).

An example of one of the larger gaps: both the 2005 and 2008 guidelines were centered around MODS application profiles, with the 2008 guidelines including flat, key-value paired Dublin Core and a thesis schema only for crosswalking to meet the Networked Digital Library of Theses and Dissertations (NDLTD) ETD-MS exchange standard.[3] Yet, as Vireo has developed, it has proliferated the generation of flat, key-value-paired metadata without equivalent attention to MODS. Several members of the working group observed that, at some point in the past five years, ETDs exported from Vireo into their DSpace repositories stopped including MODS files at all. TDL staff have indicated that a longstanding error in Vireo has blocked the production of these MODS files. The working group, weighing the concurrency problems posed by MODS files in DSpace as well as our own time and expertise, decided against providing updated MODS mappings in these Guidelines.[4]

Other gaps are anticipated rather than current: how will institutions differentiate between students, advisors, and committee members who have the same names, and how will efforts like ORCID help solve these disambiguation problems? How can we better note, in metadata, who holds the copyright to these works, whether a Creative Commons license has been applied, and whether supplemental files or data are included here, or elsewhere? Since their inception in the 1990s, ETDs have been championed as a way for students to produce more expressive work, unbounded by the constraints of the printed manuscript. How can awareness of attendant metadata issues shepherd this ambition into reality, and help users locate and discover these works?

Charged with updating the Texas Digital Library's guidelines for descriptive metadata for electronic theses and dissertations, last revised in 2008, our working group faced a number of questions and judgment calls. We have aimed to produce platform-agnostic guidelines, applicable whether institutions use DSpace, Vireo, Fedora, Digital Commons, or any other repository or thesis submission system. In crafting these updated guidelines, we have preferenced high semantic interoperability. But we also acknowledge an imperative to be practical, to produce guidelines that might be applied without introducing too great a gap between new and legacy metadata. A radically divergent standard, however closer it might be to a robust ideal, is unlikely to be successfully implemented by pragmatic institutions, or to be interoperable between systems. We hope that the guidelines you see before you successfully balance these objectives. Subsequent work, produced in concert with updates to Vireo that implement these guidelines, will provide clear guidance in approaching legacy metadata issues. Our methodology for evaluating and analyzing ETD metadata standards and practices revealed a variety of both; and we anticipate that these guidelines, rather than providing a panacea, will interact with these existing practices and standards, as institutions face trade offs in blending local application with an array of standards, tools, and platforms.

---

[3] See section 5 of this document for more information about crosswalking and sharing ETD metadata according to NDLTD standards. Networked Digital Library of Theses and Dissertations, "ETD-MS v1.1: An interoperability metadata standard for electronic theses and dissertations," ed. Thom Hickey, Ana Pavani, and Hussein Suleman. *http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html*

[4] See "A note on MODS" in the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2* for further discussion of this decision.

The recommendations included here address the current gap and consider emergent issues, suggesting that their resolution will require a combination of changes to local and community practices and to Vireo itself. Chartered around descriptive metadata needs, this working group has at times been tempted to address the many needs around preservation metadata; we hope that subsequent efforts will be formed to do so in a more concerted manner, taking up a growing movement around lifecycle management of these resources.[5] While these guidelines aim to be platform-agnostic, we provide specific guidance around Vireo and DSpace implementation. Future groups may adapt these guidelines more explicitly to Fedora-based repositories. See the "On the horizon" section of this report for more information about these issues.

Further, the TDL metadata community and the Vireo User Group are committed to work together to avoid future lags or gaps between formal/community standards and the de facto standard represented in the metadata actually generated in Vireo, metadata necessary to ensure the long-term stewardship of these important and unique student works.

[5] Daniel Alemneh, et al, *Guidance Documents for Lifecycle Management of ETDs* (Atlanta: Educopia Institute, March 2014, v. 1.0).

Contributors

- Working Group members
  - Sarah Potvin, Chair (Texas A&M University)
  - Kara Long (Baylor University)
  - Colleen Lyon (University of Texas)
  - Kristi Park (Texas Digital Library)
  - Monica Rivero (Rice University)
  - Santi Thompson (University of Houston)

- Task & Review Group members
  - Daniel Alemneh (University of North Texas)
  - Shelley Barba (Texas Tech University)
  - Christine Brown (Texas A&M University)
  - Melanie Cofield (University of Texas)
  - James Creel (Texas A&M University)
  - Thomas Dowling (Wake Forest University)
  - Jeremy Huff (Texas A&M University)
  - Sevim McCutcheon (Kent State University)
  - Billie Peterson-Lugo (Baylor University)
  - Amy Rushing (University of Texas at San Antonio)
  - Ben Steinberg (Harvard University)

- Vireo Steering Group members who reviewed the recommendations
  - Stephanie Larrison (Texas State University)
  - Joy Perrin (Texas Tech University)
  - David Reynolds (Johns Hopkins University)
  - Laura Spradlin (University of Illinois at Urbana-Champaign)

Acknowledgments:
The working group would like to extend our sincere gratitude to members of the Task & Review group and the Vireo Steering Group, who provided valuable feedback on these Guidelines. We are additionally thankful for feedback from Ryan Steans (Texas Digital Library), attendees at the 2015 TxETDA / USETDA Region 3 conference at Baylor University, and those who participated in informal interviews that helped shape the direction of the effort. Finally, we extend special thanks to Amy Rushing (University of Texas at San Antonio), chair of the 2008 TDL metadata working group, whose observations provided the impetus for our work.

Methodology

The TDL ETD Metadata working group updated the ETD descriptive metadata standards in three phases: data gathering; case study development; and outreach and results sharing.

The working group first assembled research relevant to ETD standards and practices. Through this data gathering, we identified gaps and trends that would be addressed in the updated standard. To collect data, the working group divided into two teams. One team reviewed existing ETD metadata standards, guidelines, and publications to compile common practices and reported needs among the larger ETD community. A second team conducted ten informal conversations with TDL ETD stakeholders, located within both graduate school offices and libraries, to better understand their current practices, their reliance on and usage of the current TDL ETD Metadata Guidelines, and their opinions on what topics should be addressed in a new standard. Once both tasks were completed, the two teams documented their results in a spreadsheet. They finalized this document by merging similar topics, eliminating topics that fell outside the scope of the working group, and tabling a smaller number of topics for work to be completed by a future group or task force. In the end, the working group identified five case studies: name disambiguation, availability/access metadata, complex objects and supplementary files, dates, and crosswalking.

Phase two focused on the development of these five case studies, with a goal of establishing recommendations that would become part of the updated TDL Descriptive Metadata Standard. Again, the working group divided into teams, one per case study, to complete this work. While each team created individualized processes for formulating final recommendations, many followed similar procedures. Teams conducted literature reviews related to their specific case study; they completed environmental scans to compare metadata practices and records produced at NDLTD institutions; some conducted OAI harvests of ETD metadata from open repositories to better compare how standards were implemented across institutions. After completing this process, the teams drafted summaries of the case studies, including the rationale, the issues that led to addressing the topic, and recommendations for addressing the issues.

Throughout the process, the working group shared their progress with and sought feedback from ETD stakeholders. In February 2015, the working group presented three early case studies to the TxETDA/USETDA Region 3 Joint Conference. In September 2015, the working group submitted a complete draft of their findings with the group's Task and Review Group, a team of expert stakeholders who provided input and feedback on final recommendations. Additionally, working group member Colleen Lyon served as the liaison to the Vireo Users Group (VUG), which ensured that developments were communicated to VUG leadership. This report incorporates input from these sources.

Case Studies and Recommendations

The following recommendations are centered around the five case studies: (1) name disambiguation, (2) availability and access, (3) complex objects and supplementary files, (4) dates, and (5) crosswalking ETD metadata across different metadata standards. Each case study section provides a brief summary and recommendations. The research into each case study yielded extensive findings, which have been incorporated into the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2*; see the *Dictionary* for a summary of differences between versions 1 (2008) and 2 (2015) of the Guidelines .

Name disambiguation

Institutional repository administrators face complex challenges when seeking to create consistent identification of authors and contributors in their IR. The working group found different approaches to name disambiguation, developed out of practical considerations for expedient ETD processing. The variety of approaches we observed reflect differing local policies, levels of staffing, and institutional needs.

There are a range of workflows around name disambiguation that range from higher to lower "touch," meaning a staff member is spending more or less time working with or "touching" an individual ETD record. A "high-touch" process may resemble the workflow at the University of Texas at Austin, where librarians are doing name authority work using NACO guidelines for names of faculty advisors. A "low-touch" process may involve reviewing the author and advisor names for accuracy, but making no changes to the entry to establish the names as unique to the institutional repository. Many institutions fall somewhere in between these practices.[6] So, while much of the processing of ETDs has become automated, name disambiguation is still done "by hand," when it is done at all.[7]

ETD authors generate metadata at the time of the submission of their document. Graduate school staff, librarians, and repository administrators may be far removed from the point of submission once review of the ETD record begins. It is most efficient to capture correct, unambiguous, identifying information about the author, the ETD advisor, and other contributors at the point of submission. Correcting,

[6] We observed that some institutions have trained catalogers and other staff processing ETDs to format personal names in accordance with NACO guidelines, but are not, in fact, creating authority records through NACO, nor have they received training from NACO. This approach can provide a certain measure of name disambiguation within the repository, but these name headings may not be unique outside of the local environment.

[7] This statement was first made by Boock and Kunda in 2009, but our conversations with catalogers and other university staff who work with ETDs revealed that this was still the case. See: Michael Boock and Sue Kunda, "Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries," *Cataloging & Classification Quarterly*, 47:3-4 (2009): 297-308. http://dx.doi.org/10.1080/01639370902737323

revising, or replacing name entries following submission or even ingest of the item is labor intensive, and may not be successful if the student has not provided enough identifying information at the time of submission.

Recommendations for Institutions

- Objective: Capture and display unique forms of names to enhance search and discovery in ETD collections.
    - Determine if there is a controlled name heading for the ETD author or faculty advisor, and use this form of the name in the ETD record.[8]
    - Even with the greater adoption of author profiles and researcher identity management tools such as ORCID and ResearcherID, traditional library methods of name disambiguation still result in the high quality metadata for personal names in ETD records.[9] NACO standards for name disambiguation have been widely adopted and implemented in academic libraries.[10] Using authority-controlled names in ETD records creates a correspondence between the ETD record and other published works by both ETD authors and faculty advisors. Exclusively using authority-controlled headings for name disambiguation and formatting may require a "high-touch" method processing and may not suit the needs of many institutions. A combination approach that includes the natural-language reading of a personal name alongside a unique identifier would provide the best control over author and advisor names.
- Objective: Efficiently create (and use) unambiguous names for faculty advisors.
    - Create and maintain a local authority file of frequently used and cited names in the ETD collection.
    - The use of a local authority file of frequently used and cited names in the ETD collection could help mitigate the time consuming nature of name disambiguation.[11]
- Objective: Collect sufficient information about an ETD author or faculty advisor so that a unique form of a name can be entered into the ETD record.
    - Capture as much *unique* information about an author as possible. It may not be necessary to use all of the gathered information to create a unique form of the author's name, but collecting the information in Vireo will prevent needing to solicit the author for additional information during the cataloging and archiving process.

[8] The Virtual International Authority File (VIAF) is an OCLC-hosted service, aggregating name authority files from around the world, including the Library of Congress. VIAF is openly accessible on the web; see: http://viaf.org/

[9] For more information about ORCID; see: http://orcid.org/. For more information about ResearcherID; see: http://www.researcherid.com/

[10] NACO is the Name Authority Cooperative Program of the Program for Cooperative Cataloging for the Library of Congress, through which partner libraries contribute authority records for personal names and other types of authorized headings. See: http://www.loc.gov/aba/pcc/naco/

[11] The Vireo Users Group is considering developing functionality for Vireo to interact with and check against a locally maintained file of authorized names or terms. This feature would have applications for authorized names as well as for local controlled vocabularies for department names, keywords, etc. The ability to link Vireo against an external authority files, such as the VIAF, requires further investigation.

- Middle names, middle initials, and birth years are useful in establishing a unique heading for an author or advisor.

On some campuses, students and faculty have expressed privacy concerns regarding the collection and potential display of their birth year as part of their ETD metadata. Local policy addressing these concerns may limit or eliminate the use of birth year as a tool for name disambiguation in the IR.

ORCID, and other digital identifiers, have the potential to disambiguate one author from another while also maintaining the privacy of the student or advisor. While ORCIDs were not expressly designed to replicate or replace traditional methods of name authority control, they can - as unique identifiers - functionally disambiguate name headings in ETD collections. ORCIDS can also make clear connections between names that are already highly controlled through traditional means but are still ambiguous. For example, advisors with the same name who are in different departments can be further differentiated by an ORCID. Conversely, an advisor or author who has submitted or advised ETDs in dissimilar or highly interdisciplinary fields may not be easy to distinguish by departmental association or field. An ORCID would effectively link that diverse work to the same person or persons. At present Vireo accepts ORCIDs as part of the student-submitted metadata, but best practices around the use of ORCID in Vireo have not been established. Please see Appendix E of the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Dissertations and Theses v. 2* for more information about ORCID.

Availability/Access metadata

Users have expressed confusion over rights issues for ETDs, including who owns the copyright, the terms of the copyright, and whether the item is accessible. In an environmental scan, we observed some of this information being embedded into ETD documents themselves, or reported on collection-level web pages. Occasionally, it is included directly in item-level metadata, in a number of different fields. Given the clear value in recording information about rights as item-level metadata, and doing so consistently, the working group's goal is to make recommendations updating the current standard for how access and copyright statements are made explicit in metadata, to be interpreted by machine aggregators as well as human readers.

Prior TDL metadata guidelines did not address metadata related to access, rights, or re-use status of ETDs. NDLTD's ETD-MS recommends the inclusion of an optional, repeatable dc.rights field, to be populated with "Information about rights held in and over the resource. Typically, this describes the conditions under which the work may be distributed, reproduced, how these conditions may change over time, and whom to contact regarding the copyright of the work." ETD-MS specifies three "levels: 0 (not publicly accessible); 1 (limited public access); 2 (publicly accessible)."[12] The working group also examined more robust PREMIS metadata standards and practices in communities focused on open

[12] NDLTD, "ETD-MS v1.1: An interoperability metadata standard for electronic theses and dissertations."

access and data management. The working group focused on flat metadata fields as the basis for our recommendation.

Our environmental scan revealed the development of local fields and practices for managing embargoes in repository environments. Some institutions maintain embargoed ETDs in Vireo, and others include them in DSpace as restricted files. Institutions that maintain embargoed ETDs in DSpace have developed metadata practices aimed less at description and more at functionality around releasing and managing these embargoes. To that end, the Dictionary includes examples of local fields (see section "Embargo" in the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2*) that schools have used to control the functionality of embargoes within DSpace rather than dictating any requirements. We do, however, recommend the application of human-readable language to designate availability, including indicating when embargoes are set to expire.

Recommendations for Institutions

- Objective: Communicate the rights and re-use status of the ETD.
    - Recommendation: The dc.rights and dc.rights.rightsHolder fields of the ETD metadata should be populated with statements regarding the rights and re-use of the work. This could take shape as a "canned" statement supplied from the institution to populate dc.rights. A dc.rights.rightsHolder field should communicate specific copyright ownership and data (e.g., "Copyright 2013 © Janelle Williams").
- Objective: Communicate the availability of the ETD.
    - Recommendation: A metadata field (dc.rights.accessRights) that will communicate, in human-readable language, the availability of the document (e.g,. "Publicly accessible" or "Access restricted due to embargo. Release date MM-DD-YYYY"). How this information would be maintained and updated in the repository is best determined by the institution.
- Objective: Communicate the application of a Creative Commons license or licenses.
    - Recommendation: For institutions that want to offer this option, this information should appear in dc.rights and dc.rights.uri field, with a URL linking to the appropriate license.
    - Recommendation for Vireo: We recommend a functionality in Vireo that would allow the student to select a Creative Commons license for their work, and which would then automatically populate dc.rights.uri.

Complex Objects and Supplementary files

ETDs and the research used to create them are becoming more dynamic and complex as scholarship evolves with advances in technology. Students in a variety of disciplines are beginning to create ETDs that transcends the static format of a PDF. Others are actively generating data sets and visualizations, multimedia objects, geographic diagrams, and other content in the course of their research and are increasingly interested in making this data available alongside their ETD. As demands for access to these new types of ETDs and supplemental and dynamic content increases, the need for metadata to describe

and manage this content is vital for long-term access and preservation. We recommend early actions that institutions can take to confront this emerging issue.

To identify potential recommendations, members of the working group conducted an environmental scan of peer institutions. Working group members browsed ETDs by Type or, if this option was unavailable, reviewed Format information located in an object's OAI METS record to identify examples of ETD records with complex or supplemental files. We compared common practices among the institutions to establish these initial suggestions for describing supplemental files and complex objects.

Recommendations for Institutions

- Objective: Clearly identify ETD records with supplementary or complex files.
    - Assign a type to all files uploaded to Vireo, as a way of identifying supplementary and complex files.
    - This could be done in Vireo with the use of a file characterization utility to analyze uploaded files and assign or suggest a value to be displayed in either dc.format.mimetype or dc.type.dcmi fields.
- Objective: Provide appropriate contextual information regarding supplementary or complex files.
    - Support inclusion of readme files to aid in interpreting supplementary files. We recommend that Vireo be able to ingest readme files and deposit to repository.
- Objective: Allow ETD authors to more easily include complex or supplementary files that cannot otherwise be included with the ETD record.
    - Allow students to link to supplementary materials that have been deposited elsewhere.
    - Create a prompt in Vireo for the student to provide a link to materials in Figshare, Dataverse, or other data repositories. This link would then be shared in the metadata as supplemental materials.

There should be a clear designation of supplementary files. Supplementary files could be more clearly designated with a "canned" text to appear alongside non-primary bitstreams in DSpace output. Please see Appendix D in *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2* for specific metadata recommendations regarding supplementary and complex files.

Dates

ETD records exported from Vireo include several dates, across many fields; as the current metadata exported from Vireo diverges from the 2008 TDL guidelines, these dates are particularly hard to interpret and understand. Our goal is to provide clarity around the inclusion of crucial, meaningful dates, to enable semantic interoperability, and to expand the guidelines to include information, where applicable in the standard, on the many administrative dates now collected by Vireo. Our approach here is both forensic-- what are these dates? what do they mean?-- and prescriptive.

Recommendations for Institutions

- Objective: Account for all the dates generated either by the user, ETD administrators, or the systems that manage these documents, both administrative and descriptive. Prioritize inclusion of "most important" dates, as deemed by stakeholders and evidenced in environmental scans.
    - Recommendation: Graduation Date and Date Made Public appear to be the most important to stakeholders. The placement of these two values will differ from the 2008 guidelines, as well as current Vireo implementation (see Vireo recommended section for our recommendations). Graduation date should be encoded in dc.date.issued and may be provided in natural language form in dc.date.created. The date the ETD is made available should be encoded in dc.date.available. For items under embargo, dc.date.available should coincide with the end of the embargo, when the ETD is publicly available.
    - Recommendation: Other date fields will be revised and enhanced with increased reliance on provenance fields to supply additional context for ambiguous date values. Given the likelihood of fields to change meaning over time, explicit encoding of meaningful lifecycle dates in dc.description.provenance fields will help administrators make sense of the myriad dates associated with items. See the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2* for full recommendations. Revisions significant enough to impact Vireo functionality are identified in the "Recommended Changes to Vireo" section in this document.

Crosswalking

Fewer libraries are dedicating the necessary time and effort into creating or transforming ETD records into MARC records for display in the library catalog. Instead, institutions are relying on web-scale discovery services to aggregate records across library platforms for search and discovery.[13] In a library environment that includes web-scale discovery, ETD records should be ready for crosswalk and display outside of their native systems from the point of ingestion with little to no intervention by staff. Most institutions have implemented the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to share ETD metadata across collections platforms as well as with content aggregators outside of their institution. Lack of consistency in ETD metadata, however, can result in mismatched fields, partial harvest of records, and other inaccuracies that affect end-user search and discovery. Also, given the dominance of search engines in academic search behavior, providing bibliographic data in a recommended manner may help rankings in search indices such as Google Scholar[14]. To build greater awareness around best practices for aggregation of metadata and improving discoverability, crosswalks

[13] In some instances, creating MARC records in addition to records in the IR can result in duplicate records appearing in a web-scale discovery system, which is pulling in records both from the library catalog and the institutional repository.

[14] Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. https://jira.duraspace.org/secure/attachment/13020/Invisible_institutional.pdf

from qualified Dublin Core to OAI (OAI-ETD) and to Google Scholar's Highwire Press tags have been provided as part of the TDL ETD metadata standard.

Through a series of OAI harvests, we discovered inconsistencies in mapping. We harvested ETD metadata from the institutional repositories of seventeen colleges and universities, including schools based in the U.S. and internationally. We organized our findings according to school and field, allowing us to identify trends and patterns. More than half of the institutions queried did not expose degree information in the harvestable metadata. The current crosswalk relies on a deprecated element (dc.contributor.author) to expose ETD authors, while the current standard (ETD-MS v1.1) recommends the use of dc.creator. Multiple dates are expressed in the harvestable metadata, including system-generated dates. All of the institutions we harvested from exposed between two and six dates for every single ETD record. This is confusing, and best practices should be established and implemented to limit the exposure of superfluous date fields that do not aid in search and discovery. The issues surrounding dates in ETD records are certainly complex, and warranted investigation in their own right (see above).

Recommendations for Institutions:

- Objective: Provide reliable, quality ETD metadata accessible through OAI harvesting.
    - The OAI-ETD format is not up to date with the latest version of the NDLTD's standard (ETD-MS v1.1). Our recommendations for updates to the crosswalk can be found in Appendix C of the *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2*.

Vireo recommendations

The recommendations of this group have implications for the functionality of the Vireo software. We hope that these requests can be considered during future development cycles so that Vireo functionality aligns closely with best practices in the metadata community. We detail each Vireo functionality request below.

Name disambiguation:

- Vireo should have the ability to integrate and reference a locally controlled authority file that could be used to check author names, faculty advisors, and locally controlled subject terms. Institutions should have the ability to turn this function on or off and to select their own mapping.
- Vireo should export ORCID data from Vireo using dc.identifier.orcid element.

Availability and access metadata:

- In order to provide more comprehensive information about rights, Vireo should allow for the population of a dc.rights field with a rights statement that can be customized at the institutional level. We recommend that this field include the name of the rightsholder and date of copyright (e.g. Copyright © 2015, Kristi Park).
- Vireo should provide an automated way to populate dc.rights.accessRights to indicate the availability of an ETD (e.g. Publicly accessible or Access restricted due to embargo). Institutions should have the ability to turn this function on or off and to select their own mapping (example: UT-Austin uses dc.rights.restricted for this information).
- Vireo should incorporate the ability for students to select a Creative Commons license for their work and the associated automatic population of dc.rights and dc.rights.uri fields with the CC information. Institutions should have the ability to turn this function on or off.
- Map embargo lift dates collected in Vireo to dc.rights.accessRights with standard language (e.g., “This item is currently under embargo and will be made publicly available on YYYY-MM-DD.” Or: “Access restricted due to embargo. Release date MM-DD-YYYY”). System functionality should assure that the date that appears here corresponds with dc.date.available (assigned by DSpace) and dc.date.issued. Those institutions who maintain embargoes in DSpace will need to determine a plan for maintaining and updating these values if and when embargoes are extended. How this information would be maintained and updated in the repository is best determined by the institution.

Complex objects and supplementary files:

- Vireo should analyze uploaded files and assign/recommend either a dc.type.genre value or a dc.type.dcmi value.
- Vireo should include functionality that would prompt and enable the inclusion of a readme file as a supplementary file to the ETD.

- Vireo should automatically generate text that identifies supplementary files so they aren't confused with the primary files in the public repository (e.g. add text "Supplemental document" to bitstream description.)
- Vireo should allow for the inclusion of links to data in FigShare, Dataverse, or other data repositories. That link would then become a piece of metadata associated with the ETD, such as in the dc.description.

Dates:

- Discontinue the mapping of approval date to dc.date.issued. Instead, configure this field to map graduation date in YYYY-MM format.
- Graduation date may also be mapped to dc.date.created using natural language (e.g. May 2015).
- Discontinue use of dc.date.submitted.
- Graduation date should also map to dc.description.provenance with standard language (e.g. student, FirstName LastName, graduated on YYYY-MM).
- Continue and extend the practice of generating descriptive statements around lifecycle management, and mapping to dc.description.provenance.

Type:

To align with standard, the following minor changes should be implemented:

- Recommend "Text" (rather than "text") be mapped to dc.type.dcmi instead of dc.type.material.[15]Recommend "Thesis" be mapped to dc.type.genre, per 2008 standard.

Export/mapping of degree metadata:

There is much confusion about the mapping of the following degree information fields. To make these fields more transparent, we recommend updating the Vireo mapping to export all fields as shown in table below. (This change will not impact how individual institutions may define their lists of programs, departments, or colleges, as such information is unique to each university.)

We also request that any blank values not be exported. For example, in cases where a college does not have departments, a "none" value selected in the Vireo form should not be exported as "none" to the repository but left blank.

Vireo Field	DSpace Metadata element
Program (drop down menu)	<thesis.degree.discipline>
Department (drop down menu)	<thesis.degree.department>

[15] DMCI is the controlled vocabulary recommended and general practice is to use the vocabulary as the qualifier (e.g. subject.lcsh). "Material" is more in line with "genre" (see example at EThOS UKETD_DC application profile). Also use first letter capitalized form of the term, a slight shift from 2008 standard ("text").

College (drop down menu)	<thesis.degree.college>*
Major (free text, multiple values allowed)	<thesis.degree.major>*

* New qualifiers to be added to thesis schema in the metadata registry.

General improvements to support the continuing development of Vireo metadata:

- Given the benefit of including controlled vocabulary lists for several fields (as outlined in these Guidelines), and growing opportunities around linked data, we recommend that, over the long term, Vireo functionality should expand beyond locally-maintained static controlled terms lists and include the ability to dynamically point to external controlled vocabularies. This would enable term lists to be updated continually, without local intervention.

On the horizon: Emerging case studies and next steps

A number of emergent case studies and concerns presented themselves over the course of research to inform our update to the TDL ETD Metadata Guidelines: these cases were suggested by broader trends, developments, and efforts in both ETD and metadata communities. While beyond the scope of the recommendations and updates included here, these issues might be taken up by subsequent efforts (or might be subsumed by other, not-yet-emergent issues).

1. *Linked Data*

   Every member of the working group-- and all known Vireo users-- come from institutions that use Vireo in conjunction with a DSpace repository. Our recommendations were necessarily constrained by DSpace: its data model, automated or manual processes for assigning metadata, approach to versioning, and display capabilities. While our recommendations were being drafted, DSpace 5 was released, providing support for publishing DSpace objects as linked open data.[16] This new functionality requires the application of a package and plugin to convert existing DSpace content into RDF and the installation of a triple store to store and expose the RDF (via a SPARQL endpoint). This development is exciting, but untested by working group members. As we upgrade to DSpace 5, we are interested in expanding ETD metadata recommendations to incorporate linked data guidelines.

2. *Fedora*

   The authors of these recommendations, like those that produced the 2005 and 2008 guidelines, worked towards an ideal of repository-neutral guidelines. But, as mentioned above, the constraints of DSpace, and its dominance in the TDL and Vireo User communities, provided an argument for tailoring some recommendations to the known constraints and behavior of DSpace repositories. As Vireo and TDL diversify to incorporate Fedora repositories, greater awareness should be paid to the aspects of the guidelines that are not repository-neutral, and to considering the need to tailor recommendations to Fedora and other repository systems.

3. *Preservation*

   The TDL ETD metadata working group was chartered around the consideration and update of descriptive metadata. There is, however, a pressing need for a closer consideration of preservation needs (whether related to metadata or not). Some of these concerns are likely to be addressed by working groups dedicated to TDL's emergent preservation solutions rather than in the confines of an ETD effort. Ideally, those recommendations could be reincorporated into this metadata work, in an effort to more holistically guide ETD metadata practices that are necessarily hybrid, incorporating descriptive, technical, administrative, structural, and preserving metadata in the course of stewarding ETDs through their lifecycles.

[16] See DSpace 5.x Documentation, "Linked (Open) Data," https://wiki.duraspace.org/display/DSDOC5x/Linked+(Open)+Data

4. *Beyond the union catalog*

   Due perhaps to the enduring influence of the Networked Digital Library of Theses and Dissertations, as well as to a longer tradition of cataloging print resources and more recent trends around aggregated and federated repository content (the Digital Public Library of America, Open Access Theses and Dissertations, etc.), ETD metadata efforts have tended to coalesce around a union catalog model. Rather than emphasizing the potential of full text indexing or automatic metadata assignment, standards (with some notable exceptions, the French Thèses Électroniques Françaises (TEF) standard chief among them) have focused around the creation and assignment of consistent, basic descriptive metadata. As the above discussion around preservation indicates, we are interested in the emergence of a more holistic approach to ETD metadata that goes beyond simple description, and in working towards practices and tools that fulfill early ambitions of the TDL metadata efforts. We hope our recommendations around the incorporation of stronger rights and date metadata-- as well as author identifiers-- is a first step in this direction.

5. *Beyond the PDF*

   Pioneers in the early ETD movement emphasized the expressive capacity of the electronic medium. No longer bound by what could literally be bound-- by the constraints of the printed page-- theses could take more creative, expressive forms. And yet the constraints of our systems and tools-- and regulations-- have continued to enforce the production of the electronic equivalent of that bound, printed edition: the PDF. Throughout, sound recordings, drawings, data, and other digital media have found their way into these works: whether tucked into the pocket of a printed work, folded into a figure (or, more literally, produced as a fold-out), or submitted as a supplementary file. There are signs of a resurgence towards that expressive intention and a loosening of the constraining systems and tools, and, with it, we will see metadata needs: the ability to describe and steward data files, to denote or preserve dynamic websites, etc.

Works Consulted

Alemneh, Daniel, et al, Guidance Documents for Lifecycle Management of ETDs (Atlanta: Educopia Institute, March 2014, v. 1.0). http://educopia.org/publications/gdlmetd

Arlitsch, Kenning and Patrick S. O'Brien. (2012). "Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar." *Library Hi Tech* 30 (1): 60–81. doi:10.1108/07378831211213210.

Beall, Jeffrey. (2010). Metadata for Name Disambiguation and Collocation. *Future Internet* 2:1-15. doi:10.3390/fi2010001

Bolikowski, L., Dendek, P.J. (2011). Towards a flexible author name disambiguation framework. http://depot.ceon.pl/handle/123456789/73

Boock, Michael & Sue Kunda (2009): Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries. *Cataloging & Classification Quarterly* 47:3-4, 297-308.

The British Library (n.d.) "The EThOS UKETD_DC application profile." http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=The%20EThOS%20UKETD_DC%20application%20profile.

Luyten, Bram (2014). ORCIDs in the Wild: Implementing ORCIDs into Research Support and Repository Systems. Open Repositories 2014. Panel also featuring Sarah L. Shreeves, Michael C. Witt, and Urban Andersson. Recording available: https://connectpro.helsinki.fi/p38mcoi0d3r/?launcher=false&fcsContent=true&pbMode=normal

Maurer, Margaret Beecher, Sevim McCutcheon & Theda Schwing (2011): Who's Doing What? Findability and Author-Supplied ETD Metadata in the Library Catalog. *Cataloging & Classification Quarterly*, 49:4, 277-310 http://dx.doi.org/10.1080/01639374.2011.573440

Myntti, Jeremy, Cothran, Nate (2013). Authority Control in a Digital Repository: Preparing for Linked Data. *Journal of Library Metadata*, 13:95-113. http://www.tandfonline.com/doi/abs/10.1080/19386389.2013.826061#.VL5yYCvF98E

NDLTD (2010). "ETD-MS v1.1: An interoperability metadata standard for electronic theses and dissertations," ed. Thom Hickey, Ana Pavani, and Hussein Suleman. http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html

OhioLINK Database Management and Standards Committee (2014). Standards for Cataloging Electronic Theses and Dissertations-- Remote Electronic Version (non-Reproduction). Revised August 19, 2014. https://platinum.ohiolink.edu/dms/catstandards/ETD-RDA-August-2014.pdf

Salo, Dorothea (2009). Name Authority Control in Institutional Repositories. *Cataloging & Classification Quarterly*, 47:3-4, 249-261 http://dx.doi.org/10.1080/01639370902737232

Schöpfel, Joachim (2013). "Adding Value to Electronic Theses and Dissertations in Institutional Repositories." *D-Lib Magazine* 19, no. 3/4 (March 2013). doi:10.1045/march2013-schopfel.

Schwing, Theda, Sevim McCutcheon & Margaret Beecher Maurer (2012): Uniqueness Matters: Author-Supplied Keywords and LCSH in the Library Catalog. Cata*loging & Classification Quarterly* 50:8, 903-928 http://dx.doi.org/10.1080/01639374.2012.703164

Texas Digital Library Metadata Working Group (2005). "MODS Application Profile for Electronic Theses and Dissertations," Texas Digital Library. http://www.tdl.org/wp-content/uploads/2009/04/etd_mods_profile.pdf

Texas Digital Library Metadata Working Group (2008). "Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations," Texas Digital Library, June 2008. http://www.tdl.org/wp-content/uploads/2009/04/tdl-descriptive-metadata-guidelines-for-etd-v1.pdf

Thèses Électroniques Françaises (2006). "Les métadonnées des thèses électroniques françaises," second edition, Accessed August 7, 2015, www.abes.fr/abes/documents/tef/recommandation/index.html.

Walker, L.A., Armstrong, M. (2014). "I cannot tell what the dickens his name is": Name Disambiguation in Institutional Repositories. *Journal of Librarianship and Scholarly Communication* 2(2):eP1095. http://dx.doi.org/10.7710/2162-3309.1095

"Google Scholar Inclusion Guidelines for Webmasters." Accessed August 24, 2015. https://scholar.google.com/intl/en/scholar/inclusion.html#indexing