Project Status Report: Retroactive review for Electronic Theses and Dissertations

Written by Monica Rivero, Digital Curation Coordinator
Digital Scholarship Services, Fondren Library, Rice University
January, 2013

Purpose

In 2011, Rice University began the process of implementing a fully electronic submission process for thesis and dissertations (T/D) using the Vireo System [1]. Electronic submission became mandatory for all departments in 2012 concurrent with the phasing out of print versions. In anticipation of this new process and due to significant cost increases, Rice elected to not procure 3rd party scanning services of print theses and dissertations for this interim period. This left a batch of print T/D manuscripts un-scanned for the academic years 2010-2011 and part of 2011-2012. Fondren Library took on the task to scan in-house these remaining printed manuscripts.

A second component of the project is an audit of the 1990 ETDs; comparing the metadata to the actual PDF to ensure the correct file is attached to the item record. In the past, through ILL and direct e-mail requests, it has been discovered that some PDF files from year 1990 do not match the metadata associated with them. This issue has been passed on to the vendor but no correction action has been taken. This internal audit will help Fondren Library correct this issue of erroneous PDFs and help inform us on how prevalent the problem might be for possible next steps.

Metadata is a necessary part of placing ETDs online. With the onset of the new electronic submission process, a retrospective review comparing the new metadata schema (produced by the Vireo system) is necessary to ensure metadata elements used to describe earlier T/Ds are consistent with the new schema.

This report provides details on the progress to date in achieving these project goals.

Project goals:

- Scanning of print manuscripts for selected periods.
- Convert Proquest metadata to local ETD metadata schema for T/Ds scanned in-house.
- Research and document new ETD metadata schema based on changes resulting from the migration to Vireo/ETD system and make updates to historical records as needed.

Literature review

One of the responsibilities of good data stewardship is researching best practices and standards as they develop and monitoring emerging trends in the community of practitioners. By no means an exhaustive review, there are a couple of interesting articles worthy of note in a discussion of practices for ETDs (for citations, please see bibliography section at end of this report). A major theme seems to be the changing roles of librarians in moving from a print environment to an electronic system. Electronic systems are more efficient, fewer steps required to make materials available and less physical handling of hardcopy theses. Studies at the Oregan State University Libraries found significant time and cost savings between workflows of print theses to fully electronic process. This included reducing handling from 5 staff personnel to one and overall “time spent examining and collating T/Ds [were] cut in half”; and they also realized savings of thousands of dollars from not incurring binding costs (Brook 2009). Another common theme is the gaining practice of using author-supplied metadata in combination with system directed user interfaces and the declining practice of assigning Library of Congress Subject Headings to T/Ds. Research in this area has shown that student-generated metadata “is able to deliver about 90% of the record content, most of which is both accurate and findable” (Maurer 2011). A common tactic to help ensure accuracy and consistency in creating metadata is the use of online forms “with textual guidance and selective use of features (e.g., pop-up windows, drop-down menus, scrolling lists)” (Park 2009). Both of these techniques are part of the Vireo/ETD submissions system now implemented for Rice T/Ds. One study found a high correlation between applied LCSH terms and words found in papers’ title and abstract, leading the researcher to “postulate

that the indexers were perhaps overly dependent on terms in the title to determine the subject descriptors." And though early studies "from 1980 and 1960 support formal Subject terms but this is before improvements in full search searching and indexing was available" (Schwing 2012). These findings seem to strongly suggest that authors are more likely to select meaningful terms to describe their own research and along with full-text searching of the papers themselves will provide end-users with greater discoverability and thereby reduce the need for traditional subject analysis of T/D papers. One area that seems particularly suited to authority control work is name ambiguity. Names provided in T/D manuscripts (students, advisors, committee members) are usually unique to the institution and control vocabularies such as the Library of congress name authorities file (LCNAF) would not be applicable for local usage. In the case of ETDs collections, it seems a locally developed name authority list could be an approach to address this issue. Another method to help users distinguish names, is the design of user interface displays or tools such as WorldCat identities pages to help cluster variant names though there remains questions on how much information is good enough (Thomas 2011).

Scanning

Scanning work is performed by student workers. Digitization specs were based on best practices for scanning of print versions of faculty publications [2] and local testing to determine a balance between high quality and file size with local scanning equipment[1]. To ensure high quality imaging, pages were first scanned to TIFF format and then batch converted to OCR'd PDF format.

> DIGITIZATION SPECS FOR TIFF FORMAT:
> 400 ppi for pages with illustrations or photographs (24 bit depth)
> 600 ppi for textual pages, mono bit depth (black and white)

There were two scanning teams, one for bound 1990 theses (flatbed scanner) and another for loose page theses (document scanner) Scanning began in the summer of 2012. PDFs for academic years 2010-2011 and part of 2011-2012 were completed by end of 2012. Review of 1990 ETDs is fifty percent complete at time of this report and is ongoing.

> NO OF THESIS SCANS COMPLETED AS OF DEC, 2012:
> 411 Printed theses from academic years 2010-2011 and part of 2011-2012 | 100% complete
> 23 Theses reviewed from Year 1990 | 50% complete

As evidenced by the number of files produced, the scanning rate for loose-paged manuscripts was much faster than bound theses. Scanned theses from the 1990 review have been manually uploaded to the institutional repository (only the item's bitstream was replaced, no metadata changes were made at this time). Eighty-six items from the group of loose-paged manuscripts were batch uploaded in July, 2012, using basic metadata pulled from the papers directly. This second group will have metadata updates (e.g. subjects) applied in 2013 (please see section on Metadata conversion for further details).

Interim findings from 1990 ETD review

The IR has a total of 160 theses for year 1990 and 78 of these has been manually reviewed as of Dec.31, 2012. Twenty-three items were found to hold incorrect thesis papers. This is a 16% error rate. While 2 items had no bitstream associated at all, the majority held unrelated papers from University of Oregon(16) and as well as other institutions (such as Biola University, Louisiana Tech University, Seattle University, Texas A&I University and University of Tulsa)

Items selected for this review had a dc.date.issued field value of 1990 (publication year). Dates on the signature pages of each paper show dates vary between 1989 and 1990. Dates on the signature page reflect thesis defensive and acceptance.

[1] Detail scanning instructions are available on the digital project server, including settings unique to local scanning equipment.

TABLE 1: BREAKDOWN OF SCANNED THESES BY TITLE PAGE DATE

NO.	DATES FROM SIGNATURE PAGE
1	June, 1989
1	November, 1989
1	October, 1989
2	December, 1989
4	January, 1990
1	February 1990
9	April, 1990
1	March, 1990
3	May, 1990
23	**Total Theses**

Idea on expanding audit methods

A more systematic review may be an option for continued audit of the ETD collection. For example, since the signature page of a manuscript includes the name of the granting university, a script could be designed to search for key terms, such as “university” from the first few pages and extract the surrounding text. If the search result is not “Rice University”, this would flag the item for a manual review of the PDF to ascertain if the file should be replaced.

Since historically ETDs are image-only PDFs, an initial step would be to OCR the files. Batch OCR-conversion is now a fairly straightforward process using software like Adobe Acrobat or Omnipro, both available on library computers. From the OCR’d PDFs, text analysis of the title page information can be done. A concern with this approach is the quality of OCR due to poor quality source images. Many of the older PDFs are of poorer quality. And even in a high quality page image, a title page includes handwritten text (signatures) which can easily confuse the results of OCR. So this option would still be a mixed approach, requiring manual review to confirm findings, including false findings where “Rice University” does not appear in a text search due to poor OCR conversion. A suggestion may be to take a sampling (perhaps next half of 1990 papers) to test this semi-automated method for identifying possible incorrect theses. A benefit of identifying potentially incorrectly matched PDFs upfront is this figure can then be used as a basis for estimating future scanning work and associated budget needs.

Metadata conversion

In prior years, scanned T/Ds received from Proquest included MARC records which were converted to qualified Dublin core and batch ingested to the IR. Since Fondren elected not to used Proquest scanning services during this transition period, there are no MARC records for the batch ingest of in-house scanned theses. However since these last print T/Ds were sent to Proquest to be included in their database, metadata was eventually created and may be downloaded from the Proquest database for licensed users. Though items were not found in the database when scanning began, by the end of 2012 most print theses had been entered into the Proquest database and their related metadata is available for re-purpose in local IR.

Proquest generated metadata was used as a base record for local IR items and then the data converted to our customized ETD metadata schema. This process required a review of the metadata elements pushed through the new Vireo system to the IR to help ensure consistency across the collection. (Please see next section for observations regarding updates to local metadata schema).

The source metadata for T/Ds were downloaded from Proquest in either HTML (for single value fields) or simple text format (for narrative fields such as abstracts). Simple text format data was encoded as UTF-8 and converted to XML using a text editor and regular expressions. This data was converted to excel format for further editing such as adding fields needed as part of the base ETD-MS schema [3]. This included boilerplate information such

as Genre (thesis) and Material Type (text), parsing out discipline and subject keyword terms and expanding degree abbreviations (e.g. from “M.S.” to “Masters”) [2].

A final step was matching PDF files to these metadata records. Since the file naming convention for PDFs was author’s last name plus first initial of the first name (e.g. SmithJ for John Smith), this method allowed the use of filename as the key to match files to metadata. Some investigate into non-matches was necessary. Less than 8% were found not to immediately match up by filename alone and required manual review of filenames to metadata to resolve issues. There were a sundry of reasons for the non-match ups including: duplicate ids (filenames) in cases where different authors have the same last name and first initial (eg. Dong Li and Dichuan Li), same author with multiple theses (Phd and Master) ; confusion on conversion of name pairs (first name verse middle name, or multiple last names); differences in names used on thesis title page verse graduate studies list (e.g. Su, Andy instead of Su, Yue) or simple typos. Of the 30 plus investigated, only five theses were found not to have a corresponding metadata record from the dataset downloaded from Proquest in December 2012. These un-matched papers will still be ingested but with only basic level metadata and at a later point, subject terms will be added once full metadata record is available from Proquest.

In hind sight, it might have reduced some manual editing of filenames if the convention was not merely the first initial of first name but say the first 5 digits of first name (some names are too long to use the full first name, as common best practices recommend to limit filename length to 32 digits). However, this practice would not have solved some issues, such as multiple papers by same author or name variants and it might even have introduced more error into the process by having a more complex convention.

Strongly recommend independent checks to verify number of PDFs matches, such as comparing scanning tracking IDs to actual PDFs (from command line directory listing) to metadata records; this will help ensure final count is complete and correct.

TABLE 2 : CROSSWALK FROM PROQUEST TO LOCAL DUBLIN CORE[3]

PROQUEST METADATA	ETD-MS SCHEMA	NOTE
ISBN	n/a	Proquest data not used locally, data only used as key
ProQuest document ID	n/a	Proquest data not used locally, data only used as key
Publication info	thesis.degree.grantor	Boilerplate: Rice University
Title	dc.title	
Author	dc.creator	
Advisor	dc.contributor.advisor	
Degree	thesis.degree.name	Expand abbreviation (ie. Masters of Science)
	thesis.degree.level	Conditional formula based on degree name. Values are Masters or Doctoral
Identifier / keyword	thesis.degree.discipline |dc.subject	parse out discipline from other keywords
Language	dc.language.iso	Convert to ISO 3-digit lang. code
Number of pages	dc.formate.extent	
Degree date	dc.date.created	Graduation date

[2] Steps for converting Proquest data to XML/XLS format is available on Fonlibstor project server at \\ETDs\ProQuestData\Steps for converting Proquest data.docx.

[3] Please note that this is an interim crosswalk used to batch convert records from Proquest to basic ETD-MS schema. Further edits may be required to match records to final local ETD-MS schema for consistency across entire collection.

PROQUEST METADATA	ETD-MS SCHEMA	NOTE
Publication year	dc.date.issued	Publication date
Subject	dc.subject	Not always the same values as supplied in key terms, so pull in both fields
Abstract	dc.description.abstract	
	dc.type.genre	Boilerplate : thesis
	dc.type.material	Boilerplate : text
	dc.format.mimetype	Boilerplate : application/pdf
	dc.contributor.committeeMember	n/a
	thesis.degree.department	n/a (may use OGS data)
	dc.identifier.uri	System-assigned at ingest

Electronic Theses and Dissertations Metadata Schema

With the implementation of the Vireo system, ETDs are now systematically transferred to the IR on a periodic basis, the exact timing of which is dependent on administrative review by the Office of Research and Graduate Studies[4]. Descriptive metadata for electronic submissions are author-supplied. The Vireo user interface provides drop down menus and guidelines at point of data entry[1] which aids in capturing quality metadata. The Vireo system transfers metadata to the IR in accordance with the ETD-MS: An Interoperability Metadata Standard for Electronic Theses and Dissertations [3] as well as some system-related fields. The Vireo system was development by the Texas Digital Library[4] and is open source software. Detail recommendations for creating ETD metadata are available online, including mapping cross walks [5] and MODS application profile [6], a study of these documents will greatly aid in understanding the usage of qualified Dublin core elements for ETD records.

Data generated from the Vireo system is shown in the below Figure1. There are twenty-seven active elements. The first batch of Vireo generated records totaled 120 records. In figure 1, the right hand side lists all elements currently in use for Rice's ETD collection. At a glance one can easily see there is quite a bit of difference between the two schemas. Some level of retroactive cleanup seems appropriate to bring historical data to current practice, now that the Vireo system is in place (For example removal of duplicate fields like dc.thesis vs dc.degree).

[4] for more information about Rice graduate student requirements for theses and dissertations, http://graduate.rice.edu/

Figure 1: List of ETD elements and number of records associated with each (as of Jan. 23, 2012)

Elements from Vireo Generated records	120
dc.creator	120
dc.contributor.advisor	117
dc.contributor.committeeMember	119
dc.date.accessioned	120
dc.date.available	120
dc.date.created	120
dc.date.issued	120
dc.date.submitted	120
dc.date.updated	120
dc.description.abstract	120
dc.description.provenance	120
dc.embargo.lift	8
dc.embargo.terms	8
dc.format.mimetype	120
dc.identifier.citation	120
dc.identifier.slug	120
dc.identifier.uri	120
dc.language.iso	120
dc.subject	120
dc.title	120
dc.type.genre	120
dc.type.material	120
thesis.degree.department	120
thesis.degree.discipline	120
thesis.degree.grantor	120
thesis.degree.level	120
thesis.degree.name	120

All records	7559
dc.creator	7529
dc.contributor.advisor	4665
dc.contributor.author	6868
dc.contributor.committeeMember	119
dc.date.accessioned	7558
dc.date.available	7558
dc.date.created	213
dc.date.issued	7559
dc.date.submitted	120
dc.date.updated	120
dc.degree.discipline	6855
dc.degree.grantor	6856
dc.degree.name	6856
dc.description	7
dc.description.abstract	6017
dc.description.provenance	7559
dc.description.sponsorship	7
dc.embargo.lift	8
dc.embargo.terms	8
dc.format	6943
dc.format.extent	4112
dc.format.mimetype	691
dc.identifier.citation	7558
dc.identifier.digital	93
dc.identifier.isbn	5
dc.identifier.other	1
dc.identifier.slug	120
dc.identifier.uri	7559
dc.language	7436
dc.language.iso	121
dc.publisher	6852
dc.relation.ispartofseries	1
dc.subject	7465
dc.title	7559
dc.title.alternative	2
dc.type	6866
dc.type.genre	692
dc.type.material	693
thesis.degree.department	208
thesis.degree.discipline	686
thesis.degree.grantor	692
thesis.degree.level	693
thesis.degree.name	692

Observations of selected elements

Below is a list of observations and suggestions for further work. Observations mentioned here are based on a review of the ETD-MS standard [3] and TDL ETD MODS guidelines [5] in comparison to local usage. The term "Legacy Data" refers to data before the implementation of Vireo system for purposes of this report.

General benefits for this type of work include providing larger level of consistency within the collection, improving data for interoperability; support facetted searching (DSpace discovery layer) and as an initial step for Dublin core to MODs mapping (in support of a possible future software migration).

Author

Standard practice is to use dc.creator for an author's name. Local usage is inconsistent, as some records have both dc.contributor.author and dc.creator populated. Recommend global change to transfer author names from dc.contributor.author to dc.creator for consistency within local collection and adherence to standard. This change will also support citation methodology[5].

Dates

In legacy data only the dc.date.issued element (mostly expressed in simple YYYY format) is populated.

With the new system, greater distinction is made for dates. The dc.date.issued (YYYY-MM-DD) is the publication date and dc.date.created (YYYY-MM) is the date of graduation.

There is no graduation date associated with legacy data. It is unclear if this lack of graduation date may be an issue for user browsing. Retroactively populating dc.date.created would require a labor intensive review of graduate commencement information. At this time, recommend leaving dc.date.created blank for legacy data.

TABLE 3: DEFINITION OF VARIOUS DATE ELEMENTS

Field	Definition	Example (†)
dc.date	A date associated with an event in the life cycle of the resource. In the case of theses and dissertations, this should be the date that appears on the title page or equivalent of the work. Should be recorded as defined in ISO 8601 **(a)** *Date as appears on title page is typically the date a thesis has past review process. This data is not captured locally in metadata records*	FEB 2012
dc.date.created	Creation date is defined as the date the student graduates or the date the degree is conferred, expressed in YYYY-MM format. **(b)**	2012-05
dc.date.submitted	Natural language date of date.created (May or December)	May 2012
dc.date.issued	The publication date is defined as the date the ETD is released to the public. (date is automatically generated by Dspace) **(b)**	2012-**09**-05
Degree Date	For the degree date, enter the semester in which your degree will be conferred (typically your graduation semester). **(c)**	
	(†) Example http://scholarship.rice.edu/handle/1911/64606	

a) as defined by ETD-MS standard (no qualifiers specified)
b) as defined by TDL ETD MODS guidelines
c) Help tip provided on Vireo input screen

The date as it appears on the title page does not appear in any of the dates generated by Vireo system or historical used (e.g. pulled from Proquest) in Rice's ETD collection. Yet the national standard definition states this is the date that should be associated with a record. This definition conflicts with the definition provided by

[5] Chicago Manual of Style, 14th Edition, section 14.224 Thesis and Dissertations.

TDL/Vireo system. As seen in actual theses papers during this project, the dates on the title page can differ from actual graduation/publication year (see table 1). It seems this might be confusing for an end user to open a PDF with a different year than the date shown in the metadata record. However, the date on the thesis signature page may be less useful than the actual graduation date in searching. Therefore, it is recommended when looking up a particular thesis per Patron request to search years before and after the given date.

Description

Per ETD-MS standard, "dc.description is interchangeable with the qualified element dc.description.abstract" and recommends the use of dc.description.note field for "Additional information regarding the thesis or dissertation. Example: acceptance note of the department"[6]. The formal schema definition seems to proscribe the use dc.decription.note to avoid confusion with possible aggregators. However, legacy usage places additional information regarding the T/D in either an unqualified dc.description or dc.description.sponsorship. And the general local IR practice is to use simple dc description for this sort of information in other collections. Given the low usage of qualified description data (only 7 items) in Rice's ETD collection and that Vireo does not support simple dc.description element (so it's unlikely this field will be use by authors in the future), it is recommended to leave information as is and monitor community of practices.

Degree information

Both the ETD-MS standard and TDL ETD MODS guidelines recommends syntax of thesis.__ instead of dc.degree.__ (<etd:degree> is a basic MODS syntax[7]). Therefore recommend global change to transfer dc.degree.__ data to thesis.__ to support consistency within local collection and adherence to national ETD standard.

Both thesis.degree.level (Masters, Doctoral) and thesis.degree.name (Master of Science, Master of Arts, etc) are fields being populated in Vireo and may be data that can be assign retroactively to earlier records. Recommend exploring this possibility.

Another curiosity is the differences in guidelines for syntax in degree level terms:

STANDARD	TERM SYNTAX	NOTE
ETD-MS	master's, doctoral, post-doctoral	lowercase and use of apostrophe
Vireo system	Masters, Doctoral	Capitalized first letter and no apostrophe
TDL ETD MODS	Master's, Doctoral, Post-doctoral	Capitalized first letter and use of apostrophe

This ambiguity stresses the importance of consistency at the local level. Such decisions of syntax can be a bit arbitrary or depend on local systems application, so documentation of choices helps maintain consistency of metadata across the repository.

Formats

ETD-MS standard recommends the use of standard MIME types. Vireo system supplies this boilerplate data in qualified field (dc.format.mimetype) with term "application/pdf". Historically, the unqualified field was used. Recommend replacing simple dc.format field (term "PDF") with updated qualified element and term.

General IR local metadata practice is to capture MIME type automatically at the bitstream level. So capturing the MIME type at the item level is a slight deviation from other collections, but as this is a recommended practice for the larger ETD community, the community rule supersedes local typical practice in this case.

Systematic capture of MIME types at the bitstream level will however capture supplemental document types. This is an area to monitor for changes in user practice as more and various types of files may be submitted as part of a graduate's T/D work.

The TDL ETD MODS guidelines further recommends adding a term to denote if an item is born digital (<dc.format.digitalOrigin>born digital</dc.format.digitalOrigin>) or re-formatted (when item was scanned from print

[6] http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html#dc.description
[7] This is a good example of how national standards can clash with each other (not just with local application).

version). These elements could be easily added to the records (either automatically as part of the Vireo system or as part of a mediate metadata review). However, there does not seem to be a pressing need to add what are essentially boilerplate terms at this time. This is another area to monitor and whether these sorts of terms provide useful information in managing records or not.

Historically, Rice's ETD records have included a dc.format.extent element to capture page numbers. This element is not a requirement in T/D standards. Though in a digital world, the construct of a "page" is a more fluid concept, the paper length may still provide a helpful piece of information to the researcher, perhaps giving an indication of time to read the paper. There exist programs that can systematically capturing page counts of PDF documents or alternatively, we could explore requiring graduates to hard key this data during the submission process.

Publisher

Per ETD-MS standard, the definition of the dc.publisher field states "the publisher may or may not be exactly the same as thesis.degree.grantor"[8]. In case of Rice University, these field values are the same. However, local usage is inconsistent, and some records have both dc.publisher and thesis.degree.grantor populated. Recommend redacting dc.publisher field. All records should have boilerplate grantor name populated in the thesis.degree.grantor element.

Rights

The Vireo system captures a non-exclusive agreement during the thesis submission process and this is saved as a text file along with the PDF and other metadata as part of the record for that item. What is not provided as part of the record is a general rights and usages statement. The ETD-MS standard states dc.rights is an optional field and most metadata content standards recommend a general access and usage statement. It is recommended that boilerplate info be provided for each record in the dc.right element. The current default DSpace interface will automatically list dc.right data on the short item display screen, making the information immediately apparent to an end user. Also information provided in dc.right element will be accessible to any harvesters of the metadata (while the non-exclusive text file is not). However making this change will require modification to the Vireo program itself or mediated batch upload by library staff after each ingest.

Typography: Capitalization

- For qualified dc.type fields, the Vireo system using all lower case. This appears to be a carry-over practice from MARC Genre Term[9]. All other collections in the repository, capitalize the first letter of Genre and DCMI type terms.
- A large portion of legacy data for name fields and titles are entered as all caps. "Capitalization contributions cannot impact findability, although they can contribute to understandability" (Maurer,p 23). Normalizing capitalization for name fields and titles may be an area for further study. Though there are simple formulas that can batch convert case, the ambiguity of name forms and use of proper nouns in titles would require a level of human intervention beyond a quick or simple systematic conversion of case.

Other things of note

A common theme in the above suggested changes is duplication of element usage for instance dc.creator vs. dc.contributor.author or thesis.degree.discipline vs dc.degree.discipline. These earlier practices were based on best information at that time. Given the uncertainty in applying best practices when standards and or systems are changing, it is critical to monitor community practices and make adjustments locally as warranted.

Curiously, a recommended practice in the ETD-MS documentation is recording the term "unknown" if information is not available (e.g.in elements: format, abstract, degree information). This is a major disconnect with recommended practices for sharing metadata globally[10], as it creates a false positive in aggregation of large datasets. An alternative approach commonly expressed in the digital libraries community is to present such "unknown" tags to end users as part of the interface design and not in the metadata itself.

---

[8] http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html#dc.publisher
[9] http://www.loc.gov/standards/valuelist/marcgt.html
[10] "Best Practices for Shareable Metadata" Digital Library Federation (2005) http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/AppropriateMetadata

TABLE 4: SUMMARY OF METADATA ISSUES OR SPECIAL CHARACTERISTICS

ETD-MS	LEGACY FIELD	NOTE
dc.creator	dc.contributor.author	Redact usage of dc.contributor.author in ETD collection. Typography : capitalization
dc.contributor.advisor		
dc.contributor.committeeMember		New field for Vireo system
dc.subject		Key words supplied by author
dc.date.created		Date of graduation (YYYY-MM). Not available for data pre-Vireo.
dc.date.issued		Publication date
dc.date.submitted		Date of graduation (Month Year). Not available for data pre-Vireo.
dc.date.updated		System date of ingest
dc.description.abstract		
	dc.description	May switch to dc.description.note to distinguish from abstract, if usage becomes an issue with aggregators
	dc.description.sponsorship	Localized usage field (manual entry only, not part of Vireo system)
dc.format.mimetype	dc.format	Use qualified field and convert legacy data to MIME type term (boilerplate)
	dc.format.extent	Capturing number of pages is discontinued with Vireo system. Is there a method to continue this practice with born-digital materials? Do we wish to?
dc.embargo.lift		Date when embargo is to be removed
dc.embargo.terms		Date when embargo is to be removed. In theory this field may be expressed in time period (e.g. 6 months, 12 months, etc), however current setup of Vireo populates the data in ISO date format.
dc.identifier.uri		System generated at time of ingest to DSpace
	dc.identifier.digital	Local usage for only in-house scanned print T/D. Necessary to map PDFs to 3rd party generated metadata
dc.identifier.citation		Update to Chicago style citation guidelines for ETDs
dc.language.iso	dc.language	Use qualified dc field and convert legacy data to ISO 369 (B) (alpha-3 code) standard
thesis.degree.name	dc.degree.name	redact usage of dc.degree element. Consistency of terminology (Master of Science, Master of Arts, etc) Explore retroactively assign terms to all items
thesis.degree.discipline	dc.degree.discipline	redact usage of dc.degree element
thesis.degree.grantor	dc.degree.grantor dc.publisher	redact usage of dc.degree element; Data typically found in dc.publisher is populated in dc.degree.grantor (practice for ETD collection only)
thesis.degree.level	dc.degree.level	redact usage of dc.degree element; Apply controlled vocabulary terms (Masters, Doctoral) Explore retroactively assign terms to all items

ETD-MS	LEGACY FIELD	NOTE
thesis.degree.department	dc.degree.department	redact usage of dc.degree element
dc.title		Typography: capitalization
dc.type.material		Typography: capitalization
dc.type.genre		Typography: capitalization
	dc.type	Inconsistent and non-standard usage. Update records to MS-ETD standard (e.g qualified DC: dc.type.material and dc.type.genre)
dc.date.accessioned		System generated
dc.date.available		System generated
dc.description.provenance		System generated
dc.identifier.slug		System generated
	dc.right	Research practices at other institutions to determine if appropriate to add general access and use statement to all records.

Notes:

- 27 element in base ETD-MS schema
- Proposed changes will reduce 9 elements from over all set, improve consistency across all items *in the ETC collection* and reduce duplication of certain elements

Recommendations

Next steps are based on discussions with Geneva Henry, Executive Director Digital Scholarship Services and Sid Byrd, system administrator (1/25/2013).

Near Future (1-3 months)

- Update wiki documentation for local ETD-MS metadata application profile
- Compile metadata for newly scanned PDFs (including metadata updates for already ingested print T/D manuscripts for the academic year 2010-2011)
- Ingest all completed scanned PDFs
- Make updates to legacy data to match local ETD-MS schema

Immediate future (3-6 months)

- Research practices at other institutions in use of dc.rights element verse use of txt files
- For records generated from Vireo system, help coordinate systematic transfer of ETD records to Fondren catalog to support discoverability (Sid/Denis)
- Investigate automation for systematically OCR-ing all legacy PDFs
- Research identifying incorrectly matched PDFs for other years

Long term (6 months outwards)

- Investigate name authority control mechanisms available in DSpace system (r3.0)
- Investigate methods for handling name ambiguity (students, chairs, advisors, and committee members)
- Investigate if Worldcat is pulling Rice ETD data correctly after changes have been made.

Bibliography

Content standards and format guidelines

[1] Vireo ETD System: Online submission and management of electronic theses and dissertations, Texas Digital Library. http://tdl.org/etds/

Introduction to system, including screen shots of user interface

[2] Royster, Paul, "The Art of Scanning" (2011). Digital Commons / Institutional Repository Information. Paper 67. http://digitalcommons.unl.edu/ir_information/67

Detail guidelines for creating high quality scans from faculty printed papers

[3] ETD-MS: An Interoperability Metadata Standard for Electronic Theses and Dissertations, last modified June 25, 2008. http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html

[4] Harlan, Amanda, “Texas Digital Library’s (TDL) Electronic Theses and Dissertations (ETDs) Descriptive metadata Guidelines & Vireo ETD Submission System”. ALA. 2010. http://presentations.ala.org/images/3/31/Harlan.pdf

History of ETDs

[5] Descriptive Metadata Guidelines for Electronic Theses and Dissertations. Version 1.0. TDL (Texas Digital Library) Metadata Working Group. June 2008. http://www.tdl.org/wp-content/uploads/2009/04/tdl-descriptive-metadata-guidelines-for-etd-v1.pdf

see p19 Appendix A: Quick Reference Mapping Table

[6] MODS Application Profile for Electronic Theses and Dissertations Version 1. TDL (Texas Digital Library) Metadata Working Group. December 2005. http://www.tdl.org/wp-content/uploads/2009/04/etd_mods_profile.pdf.

Articles

Beall, Jeffrey (2011): Abbreviations, Full Spellings, and Searchers' Preferences, Cataloging & Classification Quarterly, 49:6, 443-456 http://dx.doi.org/10.1080/01639374.2011.595886

Boock, Michael & Sue Kunda (2009): Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries, Cataloging & Classification Quarterly, 47:3-4, 297-308 http://dx.doi.org/10.1080/01639370902737323

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

Park, Jung-Ran (2009): Metadata Quality in Digital Repositories: A Survey of the Current State of the Art, Cataloging & Classification Quarterly, 47:3-4, 213-228 http://dx.doi.org/10.1080/01639370902737240

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

Recommended reading from DSpace Release 3.0 Documentation - Authority Control of Metadata Values

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

Thomas, Bob (2011): Name Disambiguation—Learning From More User-Friendly Models, Cataloging & Classification Quarterly, 49:3, 223-232 http://dx.doi.org/10.1080/01639374.2011.560834