Creating and Evaluating Metadata for a Large Legacy Thesis Collection:
From “Vocational Agriculture” (1922) to “Microemulsion-mediated syntheses” (2004)

Background:

In the summer of 2012, Texas A&M University Libraries uploaded more than 16,000 retrospectively-digitized masters-level theses, dating from 1922 to 2004, into our DSpace institutional repository.

Item records for the Retrospective Theses collection were created by mapping existing MARC records, then transforming and enhancing this metadata. Records included fields encoded in our Qualified Dublin Core schema, as well as the custom Thesis schema developed by the TDL member consortium. MODS metadata records were also generated, to be stored as bitstreams.

More than eight decades of MARC cataloging preceded the transformation and enhancement of metadata for theses in the collection. In every sense, metadata for the Retrospective Theses was dependent on this body of MARC data.

Over eight decades, though, we see shifts in A&M’s MARC records. As Surratt and Hill (2004) observe in their review of thesis and dissertation cataloging at Texas A&M through its history, the evolution of the MARC was tied to changes in cataloging practices, including:

- Moving away from full subject analysis, owing to staffing shortages and the novelty of certain subjects addressed in student works.
- The consistent use of a local call number system rather than Library of Congress classification numbers.
- Experimentation with technology to add abstracts to bibliographic records.
- Fields used in the catalog but not submitted to OCLC.

(See Brian E. Surratt and Dustin Hill, “ETD2MARC: A semiautomated workflow for cataloging electronic theses and dissertations,” *Library Collections, Acquisitions, and Technical Services* 28, no. 2 (2004): 205-223.)

Sources of standards for ETD and TD Metadata:

Local

- Public services at A&M requested that we adhere to conventions that enabled their discovery of items, such as retaining “Major subject” for dc.subject terms.

Regional

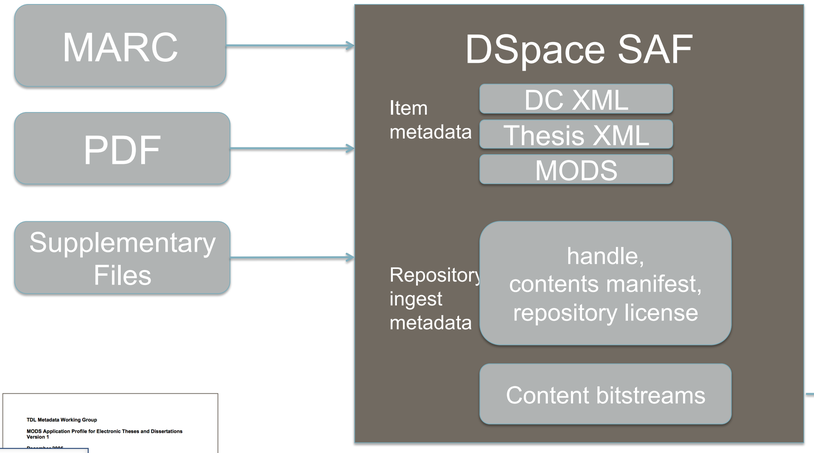
- The Texas Digital Library’s *MODS Application Profile for Electronic Theses and Dissertations* (Version 1, 2005). Established to facilitate deposit and interoperability in the shared ETD repository operated by TDL.
- The Texas Digital Library’s *Descriptive Metadata Guidelines for Electronic Theses and Dissertations* (Version 1.0, 2008). Prepared by the TDL Metadata Working Group, these standards aimed to prevent data loss when mapping from MODS to the Dublin Core standard of the DSpace TDL ETD repository.

National

- Networked Digital Library of Theses and Dissertations, *ETD-MS: An Interoperability Metadata Standard for Electronic Theses and Dissertations* (2001-2008).

Crosswalking MARC TD Metadata to DC and MODS Metadata:

Repository-suitable metadata were generated by means of three XSL transforms applied to MARC-XML from the catalog. MARC records were extracted on the basis of call-number and subjected to these transforms to produce Dublin-Core style metadata in the *dc* and *thesis* schemas as well as a MODS record. These metadata files were packaged in the DSpace Simple Archive Format (SAF) for ingestion.



Continuing challenges:

- Inconsistent (and sometimes inaccurate) metadata.
- Lack of controlled or normalized values for metadata values, including disciplines and majors.
- Additional burden of editing MODS when editing DC and Thesis DSpace metadata.
- Lessons (re)learned: lack of consistent metadata creates additional usability issues.

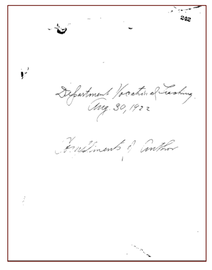
(See Adam Mikeal, Tim Brace, John Leggett, Mark McFarland, and Scott Phillips, “Developing a Common Submission System for ETDs in the Texas Digital Library,” *Proceedings of the 10th International Symposium on Electronic Theses and Dissertations*.)

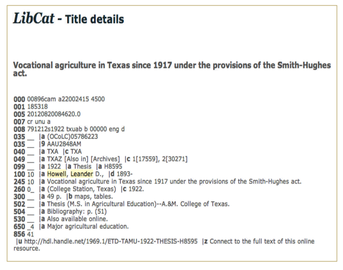
Improvements in the works:

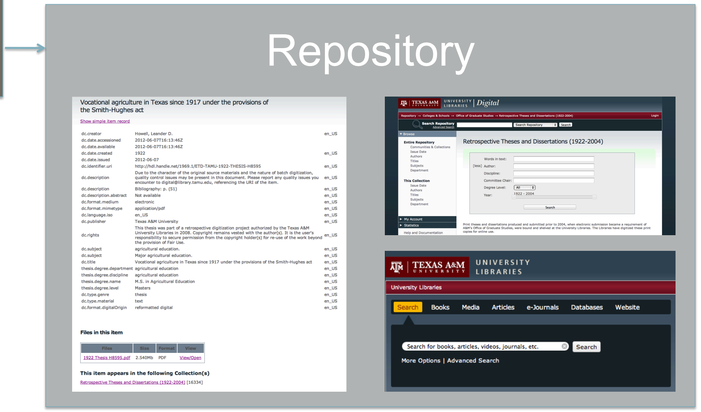
- Removing potentially inaccurate “department” values (global collection wipe of thesis.degree.department).
- Adding dc.type values that will facilitate faceted searches in Primo.
- Investigating normalizing “major subject” terms, found in four places in DSpace metadata.
- Implementing any edits in MODS.

Pie in the sky:

- Adding committee members and other metadata found in the documents themselves.
- Mapping out departmental evolution and assigning accurate department values.







Original MARC Cataloging → Evaluation of TD/ETD metadata standards → Crosswalking into DC, Thesis, and MODS schema for IR ingest → Ingest and continuing evolution of metadata

Sarah Potvin, Digital Services & Scholarly Communication, Texas A&M University Libraries
James Creel, Digital Initiatives, Texas A&M University Libraries

This project would not have been possible without the work of many in the Texas A&M University Libraries who cataloged, inventoried, packed, shipped, uploaded, mapped, ingested, assessed, and transformed the Masters Theses.

Image sources (L-to-R): page image from Leander D. Howell, “Vocational agriculture in Texas since 1917 under the provisions of the Smith-Hughes act,” Masters Thesis, Texas A&M University (1922); screen capture of Texas A&M LibCat record; mappings and title page from Texas Digital Library, “Descriptive Metadata Guidelines for Electronic Theses and Dissertations,” (June 2008); title page from Texas Digital Library Metadata Working Group, “MODS Application Profile for Electronic Theses and Dissertations,” version 1 (December 2005); screen capture of crossswalk built at Texas A&M to transform MARC records; Texas A&M Repository item record, search interface, and University Library website search interface.