

**ETD2MARC: A semi-automated workflow for cataloging electronic theses and
dissertations**

**Brian E. Surratt
Electronic Resources Cataloger
Texas A&M University Libraries
5000 TAMU
College Station, TX 77843-5000
979-845-5454
bsurratt@tamu.edu
(Please direct correspondence to this author.)**

and

**Dustin Hill
Programmer Analyst
Thesis Office
Texas A&M University
5000 TAMU
612 Sterling Evans Library
College Station, TX 77843-5000
979-845-2225
dhill@vprmail.tamu.edu**

Abstract:

This article describes a semi-automated workflow for cataloging electronic theses and dissertations (ETDs). A perl script is used to query the metadata in an institutional ETD database and create a Machine Readable Cataloging (MARC) record for each ETD. The MARC records are imported into the Online Computer Library Center (OCLC) WorldCat database using the Connexion service, proofread, updated, and exported to the local catalog. Topics discussed are the cataloging decisions that were made prior to the creation of the script, the benefits and limitations of this workflow, future applications of the workflow, and future opportunities for research.

Keywords: Electronic Theses and Dissertations, ETD, MARC, perl, Connexion

1. Introduction

Texas A&M University began a pilot program to accept electronic theses and dissertations (ETDs) in 2002. Realizing that these “born digital” documents are a permanent addition to the University’s collection, the library planned to represent them in the catalog along with print theses and dissertations. ETDs, however, present both challenges and opportunities to the established library workflow.

1.1 Cataloging print theses and dissertations at Texas A&M

The planning process for ETD workflow began with a review of the history of thesis and dissertation cataloging at Texas A&M. Cataloging policy evolved over time due to needs, constraints, and available resources. In the past, we conducted full subject analysis, but a reduction in staff required another approach. Furthermore, the subject matter of some theses and dissertations is new, thus there may be no existing subject heading to describe a given resource. We have never added Library of Congress classification numbers to theses and dissertations, preferring to use a local call number system. In the early 1990’s, we began experimenting with scanning and optical character recognition technology to add abstracts to bibliographic records.

The current policy is for staff to create a core level record. We do not conduct subject analysis, add controlled or uncontrolled subjects, or add Library of Congress call numbers. We use a locally created call number system to collocate our theses and

dissertations and arrange them by year the degree was received and author's name.

Abstracts are added to the records in the local catalog, but not to the records contributed to OCLC.

1.2 Workflow for print theses and dissertations at Texas A&M

Under current practice, the Texas A&M University Thesis Office physically delivers print theses and dissertations to the cataloging department. A library staff member manually creates a Machine Readable Cataloging Record (MARC) in the Online Computer Library Center (OCLC) WorldCat database and then exports the MARC record to the local catalog. After cataloging, the thesis or dissertation is added to the permanent collection.

1.3 Considerations for an ETD workflow

Because ETDs are digital objects that are accessed remotely over the Internet, the existing workflow for print theses and dissertations was not applicable. Furthermore, the ETD database contains descriptive metadata, which can be harvested to automate the creation of MARC records. Consequently, manual descriptive cataloging would largely be redundant for ETDs. Automating the process can potentially reduce the amount of time and effort required to "catalog" ETDs, increasing processing speed and conserving human resources that can be allocated for other uses. This article will describe the workflow that was created at Texas A&M University to represent ETDs in the catalog

and will highlight a computer program that was written to query the database and create MARC records for each ETD.

2. Literature Review

There is a small body of literature related to cataloging theses and dissertations. Hoover created a bibliography for this literature in 2001 [1]. She notes that the most common theme is the tension between the need for efficiency and the need for high quality records for theses and dissertations. Most resources in the bibliography discuss cataloging theses and dissertations in general, and were useful for making cataloging decisions. There is not much literature that discusses cataloging ETDs specifically, or automating the process, but a number of resources were pertinent to this project. In 1996, McMillan discussed the possibility that MARC records for ETDs could be created using computer programs, which would allow catalogers to concentrate on providing intellectual access to ETDs [2].

In 1999, Sharretts, Sheih, and French discussed a method implemented at the University of Virginia (UVA) for automatically creating MARC records for ETDs [3]. The authors discuss the justifications for cataloging ETDs, deciding what metadata to add to the MARC records, the coding of the MARC records, and note that automating the process results in cataloging with very little human intervention. The article describes two processes for creating MARC records. The first process used the program 'grep' to extract metadata directly from the PDF file. The resulting MARC records did not conform to

OCLC guidelines. Specifically, the program was not able to change the name on the title page to an inverted format for the main entry-personal name field (MARC variable data field 100) and could not separate the main title from the subtitle. This process was abandoned. The second process used metadata input into a web form by students rather than taking the metadata from the ETD itself. Instead of grep, the scripting language 'perl' was used to write the metadata to an ASCII MARC record. A second perl script compiled the MARC records into one file and loaded them into VIRGO (the UVA catalog). This process resolved the problem of format of the main entry. The program also searched for a colon in the title. If a colon was present, the words after the colon were placed in the other title information subfield (245 subfield 'b'). The authors point out that this is not a reliable method because students may not format the title correctly when inputting the metadata.

Highsmith, Jordan, Llona, Murray, and Summers (2002) indicated the potential for creating MARC records using an automated script [4]. Their article "MARC it your way: MARC.pm" provided examples of how the perl module MARC.pm could be used to manipulate MARC records. Further research indicated that MARC.pm was later replaced by the MARC::Record module. The documentation for this module proved invaluable in the coding of the ETD2MARC script [5].

3. The ETD system

To facilitate the management of ETDs, Texas A&M installed a modified version of the Virginia Tech ETD software [6]. This software creates a web-based application for managing ETDs in a locally maintained database. The system enables graduate students to upload ETDs into the database, provides administrative access for management of the database, and provides an archive for patron retrieval of ETDs. At Texas A&M, the system has been implemented on servers maintained by the Office of Graduate Studies (OGS), which is the parent unit of the Thesis Office. It is a homogenous, stand-alone system that is exclusively used for ETD management.

In the spring semester of 2002, the Thesis Office of Texas A&M University began a pilot program to accept ETDs. Limited academic units participated in the initial study.

Beginning with the spring semester of 2003, all graduate students at Texas A&M were provided the option of submitting theses and dissertations in electronic, rather than paper, form.

In order to submit an ETD, a student must log in to the Thesis Office web site. After logging in, he or she is presented with a web form [Figure 1]. The web form requires him or her to submit descriptive metadata about the ETD, such as the name of the author and the title of the ETD. After completing the web form, he or she uploads her ETD into the database. The ETD is then stored in an SQL database with the metadata formatted according to Dublin Core and captured in an XML wrapper. For a list of significant metadata fields, see Appendix A.

4. Cataloging workflow

4.1 Preliminary considerations

Once the ETDs are in the database, the general public can retrieve them via the World Wide Web at <http://thesis.tamu.edu>. The search engine on the thesis website only allows searching on limited fields, and does not currently allow searching on names or titles.

Texas A&M University Library decided to represent each ETD in the local catalog. Doing this achieves the objects of the library catalog, initially defined by Charles Ammi Cutter in his "Rules for a Dictionary Catalog" [7]. Furthermore, development of the objectives of the catalog is ongoing, most recently refined by the International Federation of Library Associations and Institutions (IFLA) in 1997 [8].

Each record will be made available through Texas A&M's online public access catalog, which will collocate ETDs with the print collection of all Texas A&M theses and dissertations and maintain the integrity of the collection over time. According to cooperative cataloging practice, the bibliographic record will be added to the OCLC WorldCat database, making the record available for other OCLC members to use. OCLC will also harvest these records into its Experimental Thesis Catalog (XTCat) as a part of its ETD Project [9]. Cataloging the ETDs will allow the library to verify name authorities for authors established in the National Authority File, or contribute new authority records as needed.

Manually cataloging electronic resources requires that the cataloger view the resource, conduct technical reading for cataloging purposes, and then manually code a MARC record in OCLC. A more efficient method would be to extract the metadata from the ETD database and create MARC records using a simple computer program called a “script.”

4.2 The ETD2MARC perl script

The library coordinated with the Texas A&M Thesis Office to write and implement a script for record creation. The script, titled ETD2MARC, uses the perl language and the perl module MARC::Record. MARC::Record is a set of programming instructions that can create and perform editing functions on MARC records [10]. See Appendix B for the full text of the script.

The script was written through a collaborative effort between the Library's electronic resources cataloger and the Thesis Office's programmer analyst. The e-resources cataloger wrote the portion of the script that uses commands from the MARC::Record module to create the MARC record (see Section 4 of Appendix B). Writing this portion of the script required knowledge about the structure of MARC records and cataloging rules for ETDs. The programmer wrote the rest of the script with input from the cataloger (Sections 1, 2, 3, 5, and 6). Because the ETD database resides on restricted servers owned by the Office of Graduate Studies, the programmer implemented the script into the system and performs ongoing maintenance.

The script achieves three primary functions. First, it queries the ETD database to extract the metadata for each record. Second, it creates a MARC record for each ETD, placing the appropriate metadata in the appropriate MARC tags. Finally, it saves the MARC records to a file that is accessible to library staff.

4.3 Running the script

After each ETD is submitted, it is initially stored in a database that is accessible only to Thesis Office personnel. The ETD is not available to the public until certain conditions are met, namely the student must graduate and agree to release the ETD to the public. Once these conditions are met, the Thesis Office will approve the ETD for public release. To release an ETD to the public, Thesis Office staff logs in to the administrative interface of the ETD database and changes a check box to “release to public.” This changes a binary variable that allows the ETD to display to the public on the thesis website.

ETD2MARC runs automatically when Thesis Office staff approves an ETD for public display. After the script runs, the MARC record is created and saved to a file on a library shared-access disk drive. At this time, it is accessible for library staff to upload into the WorldCat database.

4.4 Saving to WorldCat and the OPAC

To save to WorldCat, a library staff person opens a web browser to the OCLC Connexion web site [11]. He or she uploads the record to WorldCat using the import feature. This sends the MARC record to the save file, where he or she proofreads and edits the record as needed. After editing, he or she updates the record in WorldCat and exports it to the local online public access catalog, which is Endeavor's Voyager at Texas A&M.

Due to the limitations of the script, every record requires some manual editing. At a minimum, the cutter number must be added to the call number field. Second, the script cannot (at this time) determine if there are any non-filing characters in the title proper. In terms of coding the MARC record, this means that the script cannot code the second indicator of the title statement (245) tag. Third, the script cannot determine the presence of "other title information" and thus cannot code a 245 subfield 'b' if needed. The script takes all of the characters in the "title" metadata tag and transposes them to the 245 subfield 'a'. Consequently, if an ETD does have a subtitle, it is appended to the end of the subfield 'a' and must be manually moved to a subfield 'b', with any necessary corrections in punctuation. Fourth, experience has shown that many abstracts contain characters that cannot be correctly displayed in non-Unicode systems. An example is the use of Greek letters to represent mathematical symbols. In such a case, the cataloger must edit according to Library of Congress Rule Interpretation 1.0 E [12]. Finally, the cataloger must review the record for general correctness. Errors may be present due to incorrect capitalization, spelling, or other errors of data entry on the part of the graduating student.

5. Cataloging decisions

5. 1 Bibliographic characteristics of ETDs

The goal for ETD2MARC was to create core level records with OCLC encoding level 'K,' the same standard that Texas A&M meets for print theses and dissertations [13]. In order to achieve this standard, many cataloging decisions had to be made when writing the script. The first step in making cataloging decisions was to identify characteristics common to all Texas A&M ETDs.

At Texas A&M, ETDs are predominantly textual in nature and are stored in the Adobe PDF format. They are accessed through the World Wide Web, and are thus considered remote access electronic resources. In terms of issuance, ETDs are monographs. The chief source of information is the metadata that is supplied by the author. Finally, they are born digital, and not considered reproductions or versions of print resources. These characteristics determine the coding of the fixed and variable fields in the MARC record.

5. 2 Leader and fixed field

According to Anglo-American Cataloguing Rule (AACR2) 9.4B2, remote access electronic resources are considered to be published [14]. This is different than print theses and dissertations, which are considered unpublished manuscripts according to OCLC Bibliographic Formats and Standards [15]. Consequently, ETDs are “language materials” and are coded ‘a’ in the type of record position (leader/06) [16]. Being monographs, they

are coded "m" in the bibliographic level position (leader/07) [17]. The combination of these two characteristics determines the bibliographic format of the material that is being cataloged. In this case the "Books" format is used, with corresponding fixed length data elements (008) [18]. Because the type code only reflects the textual nature of ETDs, an additional material characteristics (006) field is required to describe the computer resource bibliographic characteristics of ETDs [19]. The physical description (007) field reflects the physical properties of ETDs [20]. See Appendix C for the coding of specific character positions.

5.3 Variable fields

An important step in designing the script was deciding what metadata elements would map to each MARC record. For example, the 'first_name' and 'last_name' metadata elements are mapped to the main entry-personal name (100) field. A spreadsheet was created [Appendix D] that lists relevant metadata fields from the ETD database and the MARC tags that they map to.

Each ETD MARC record includes the applicable descriptive fields, including fields to describe characteristics of theses and dissertations in general, and fields to reflect the electronic nature of ETDs. OCLC's Bibliographic Formats and Standards section 3.1 provides guidelines for cataloging theses and dissertations. AACR2 chapter nine provides guidelines for cataloging electronic resources.

The unique ETD number (assigned by the Thesis Office) is input into the system control number field (035). The MARC code for Texas A&M (TXA) is input as the original cataloging agency and the transcription agency into the cataloging source field (040). The year the ETD was accepted and the document type are input into the local free-text call number field (099). The author's name is input into the main entry-personal name field (100). The title proper and general material designation (electronic resource) are input into the title proper field (245). The year that the ETD was approved is input into the publication and distribution field (260). A quoted note from the title page indicating the major subject area of the student is input into a note field (500). A note indicating that the source of the title is author-supplied metadata is input into a note field (500). A dissertation note (502), indicating the degree granting institution is included. A note indicating the presence of bibliographical references (504) is included. A type of computer file note (516) is included. The abstract proper is input into a summary note (520). A mode of access note (538) indicating the mode of access is the World Wide Web is included. A topical subject added entry field (650) is included for the major subject area. Keywords are input into index term-uncontrolled fields (653). Finally, the uniform resource locator (URL) for the ETD is input into an electronic location and access field (856).

An item described note (500) is also required according to AACR2 9.7B22, but this note was not included in the script. This is because the rule states that the date the item was viewed should be indicated. We assumed "viewed" meant that a person looks at the resource. Because the ETDs have not been viewed at the time of MARC record creation,

this note is not created by the script, but added by hand in a later step. See Appendix E for a table that shows the coding of variable fields, including formatting, subfields, and indicators.

These cataloging decisions should not be considered authoritative for ETD cataloging. Each library must make decisions that are appropriate for its purposes. The decisions made at Texas A&M are not necessarily the same decisions that another agency might make. Furthermore, any given set of ETD metadata can be transposed to MARC records using various algorithms, so these rules need not be static. Even generally accepted cataloging standards change over time, so rules that may be considered correct today may change in the future. Although high quality cataloging based on sound bibliographic principles was one of the goals of this project, the primary goal was demonstrating that technically correct MARC records can be created by a script and input into bibliographic utilities. We anticipate that the algorithm we use to generate ETD MARC records will change in the future, for various reasons. See Appendix F for an unmodified output record.

6. Conclusions

6.1 Comparison to the University of Virginia process

The workflow implemented at Texas A&M learned from and expanded on the process described by Sharrets, et al. The projects have the common goal of creating MARC

records for ETDs using a computer program. However, Texas A&M provided a different context, and thus different challenges. The workflows differ in four ways. First, Sharrets, et al. do not address contributing the records to OCLC. This remained a priority at Texas A&M. Second, at Texas A&M, MARC records are not created for ETDs until they are displayed to the public. Some ETDs are embargoed pending publication or patent. The creation of the MARC record had to be coordinated to coincide with public release, even if the public release is after graduation. Third, Texas A&M decided to use a tool that was not implemented in the Virginia project, namely the perl module MARC::Record. Fourth, there was no intention at Texas A&M of completely automating the process. Automation was seen as an opportunity to make the process more efficient, but not completely supplant it. We still anticipated that library staff members would interact with the bibliographic record because humans necessarily perform some acts of cataloging. In our case, this mostly pertains to adding cutter numbers to the local call number and verifying name authorities. Still, the Virginia project accurately predicted many of the hurdles encountered in the Texas A&M project, such as formatting of names and titles, and was a useful resource in the development of the ETD2MARC script.

6.2 Limitations of the Texas A&M workflow

This workflow is limited by the quality of the metadata input by the student. The metadata may contain omissions, typographical errors, or entries that are not formatted according to cataloging rules. These errors must be manually corrected after uploading the record into OCLC. The occurrence of errors is somewhat mitigated by the data entry

standards of the Thesis Office. In practice, the most common errors are titles that do not conform to the capitalization rules of AACR2 Appendix A.4.

The formatting of the 245 field is a limiting factor. The script cannot determine the number of non-filing characters at the beginning of a title. The script uses a default setting of "0" (indicating no non-filing characters) for the second indicator of the 245. We anticipate that future versions of the script will address this problem. Also, the script cannot determine if a subfield "b" is needed for the presence of other title information. The majority of ETDs do not have to be further edited, but the presence of either of these cases must be corrected manually in OCLC.

The script is limited to descriptive metadata only. It is not practical to have graduate students input controlled vocabulary subject headings or controlled classification numbers at the time they submit ETDs. Therefore, certain fields that are otherwise considered valuable are not created because the necessary information is unavailable in the metadata of the ETD database. These fields are the controlled subject field (650) and the classification field (050). Furthermore, students do not verify personal author names (100 field). It is clear that descriptive cataloging can be performed by the script, but subject analysis, classification, and name authority verification can only be performed manually because they require intellectual effort beyond the capabilities of a computer script.

Limitations in utilities' character sets, as evinced by Greek characters in abstracts, present another problem. For example, OCLC's character set is ASCII based and supports limited non-Roman characters [21]. To compensate for this, the library staff must edit non-displaying characters in accordance with Library of Congress Rule Interpretation 1.0E. This will change in the future as utilities incorporate Unicode. Endeavor is supporting Unicode with its release titled "Voyager with Unicode" [22]. OCLC is upgrading the WorldCat database from a proprietary system, to the industry standard Oracle database in order to provide added features, including Unicode support [23].

The script assumes that ETDs primarily consist of text. This may not be appropriate if the main component of an ETD is non-textual, such as video, sound recording, multimedia, or other type of digital object. As ETDs evolve over time, this workflow may have to be altered or abandoned to accommodate new media formats.

The fact that the item described note is mandatory raises an interesting issue regarding traditional cataloging. AACR2 rule 9.7B22 states "For remote access resources, always give the date on which the resource was viewed for description." This rule seems to presume that a human always views a resource before it is cataloged. In modern information systems, descriptive metadata, created at a certain point in time, can be repurposed to other formats, including MARC records. The current rules in AACR2 do not account for this flexibility in the nature of modern information systems. Perhaps rule 9.7B22 should be changed to allow for the possibility that the record was automatically

created. A possible addition to the current rule could be "Alternatively, if the record was created by an automated system, give the date on which the record was created."

6.3 Benefits of the Texas A&M workflow

This method of cataloging ETDs has a number of benefits. The primary benefit is increased efficiency gained by using a script versus manually encoding MARC records. This potentially reduces the amount of human resources, and ultimately the amount of time, needed to catalog MARC records. In the past, it took months for some theses and dissertations to be cataloged. This method seeks to reduce the amount of time to days or weeks. Anecdotal evidence suggests that the time is significantly reduced, but this hypothesis has not been formally researched.

This process adds value to MARC records by including abstracts and keywords. The current workflow for print theses and dissertations does not include this information in the records contributed to OCLC.

MARC records for ETDs include a URL to the full text of the ETD, providing free access to anyone with a web browser. While MARC records for physical resources point interested users to a location on a shelf, MARC records for ETDs serve as gateways to the full-text items themselves, thus eliminating geographic location as a barrier to access.

The workflow allows ETD MARC records to be uploaded into WorldCat, thus facilitating the process of cooperative cataloging. Once in WorldCat, they are harvested into OCLC's XTCat, which increases universal access to the Texas A&M ETDs. Uploading the records into WorldCat allows the records to be proofed and edited manually if needed.

6.4 Future research and applications

Texas A&M will continue to improve this system for cataloging ETDs. The next step will be to accommodate name suffixes and numerals into the process. Students have been adding this information to the last name field, but we will modify the web form so that this is a separate field. Other potential modifications are the inclusion of committee members and department names. These data are already collected as ETD metadata, and a simple modification of the script would add this information to MARC records.

Future versions of the script will be able to determine the number of non-filing characters in the title statement for English language ETDs. All Texas A&M ETDs have been written in English so far, and it is expected that this feature will suffice for the time being. But a method for accommodating non-English ETDs will eventually have to be explored.

Future research will address whether cataloging time is saved by this process, and if so, the degree of savings. The amount of time saved is affected by many variables, such as the nature of the collection, quality of the metadata, and training of library staff.

This approach may be applicable to other digital collections with descriptive metadata. This script can query any SQL database to input metadata into MARC records. Texas A&M is exploring the possibility of using this script in conjunction with an institutional repository system. This approach would work best when a database contains objects that have consistent characteristics, i.e. a database where all the objects are one media type. At this time, the script cannot distinguish among various types of digital media, but further programming could extend the capabilities of the script.

Experimentation with this script revealed the unavoidable degree of human interaction that is required to catalog using AACR2 and MARC21. Despite having high quality metadata along with a script that can transpose that data to the appropriate MARC fields, human interaction is still required in order to meet accepted cataloging standards. An interesting question for future research might be, "Are the steps that must be performed by humans (such as counting non-filing characters, verifying authorities, and adding controlled subject headings) worth the cost when evaluated in terms of improved retrievability?"

These non-automated steps cost resources in terms of human capital. Presumably, they have a value in terms of improved data retrieval. The question to research is if the value

of improved data retrieval is greater or less than the cost in human capital. This is a compelling question considering there are alternate, non-AACR2, non-MARC21, based methods of description for digital materials, such as Dublin Core in XML or HTML. The costs and benefits of cataloging ETDs could be contrasted with ETD retrieval from non-MARC repositories.

Finally, progress will continue to be made in automating cataloging, while still meeting the standards of AACR2. Cataloging steps that can currently be performed only by a human may eventually be performed by machine. Alternate methods of retrieval, such as full text searching utilizing relevancy rankings, may one day improve to the point of negating the need for traditional description and subject analysis. The impact of automation is an ongoing theme in cataloging, classification, and retrieval research.

Acknowledgements

The authors would like to thank the following people at Texas A&M University for their assistance in the development of this project: Tommy Armstrong for providing computer systems support, Anne Highsmith for providing advice on perl scripting, Jeannette Ho for critiquing the initial ETD MARC records, Barbara McGuirk (Director of the Thesis Office) for encouraging collaboration with the University Library, Elizabeth Neff for explaining the workflow for print theses and dissertations, Laura Osborne for providing training on OCLC's Connexion service, and Rachel Stovall for assisting with the literature review. The following Library faculty generously provided feedback on the

initial draft of this article: Aletha Andrew, Lisa Furubotten, Anne Highsmith, Jeannette Ho, Jon Marner, Kathy Weimer, and Mary Dabney Wilson.

References

[1] Hoover, L. (2001). Cataloging Theses and Dissertations: An Annotated Bibliography. *Technical Services Quarterly*, 19 (3), 21-39.

[2] McMillan, G. (1996). Electronic Theses and Dissertations: Merging Perspectives. *Cataloging & Classification Quarterly*, 22 (3/4), 105-125.

[3] Sharretts, C. W., Shieh, J., and French, J. C. (1999). Electronic Theses and Dissertations at the University of Virginia. *ASIS 99: Proceedings of the 62nd ASIS Annual Meeting*, 240-255.

[4] Highsmith, A., Jordan, M., Llona, E., Murray, P. E., and Summers, E. (2002). MARC It Your Way: MARC.pm. *Information Technology and Libraries*, 21 (1), 19-25.

[5] Bearden, C., Campbell, C., Birthisell, B., Lane, D., Lester, A., McFadden, C., O'Regan, M., and Summers, E. (Sept. 16, 2003) MARC/Perl: Machine Readable Cataloging + Perl. (N.D.) Online. Retrieved from the World Wide Web: <http://marcpm.sourceforge.net>.

[6] Kletnieks, C. (Sept. 16, 2003). ETDs @ VT (ETDs at Virginia Tech). (N.D.)
Retrieved from the World Wide Web: <http://scholar.lib.vt.edu/ETD-db>.

[7] Cutter, Charles Ammi. (1904). Rules for a Dictionary Catalog (4th ed.). Washington,
DC: Government Printing Office, p. 12.

[8] Svenonius, Elaine. (2000). The Intellectual Foundation of Information Organization.
Cambridge, MA: MIT Press, pp. 15-18.

[7] Hickey, T., Young, J., and Dehn, T. (Sept. 16, 2003). Electronic Theses and
Dissertations (OCLC - Projects). (N.D.) Retrieved from the World Wide Web:
<http://www.oclc.org/research/projects/etd/>.

[8] Bearden, et. al.

[9] OCLC (Sept.16, 2003). OCLC Connexion: Integrated Cataloging Service. (1999-
2003) Retrieved from the World Wide Web: <http://connexion.oclc.org>.

[10] Library of Congress. (2002). Rule Interpretations (2002). Washington, D.C.:
Cataloging Distribution Service, Library of Congress. Retrieved from Cataloger's
Desktop (2003, Issue 2).

[11] OCLC. (2002). *Bibliographic Formats and Standards* (3rd ed.). Dublin, OH: OCLC, Inc., p. 30.

[12] American Library Association. (2002). *Anglo-American Cataloguing Rules* (Revision 2002). Chicago: American Library Association. Chap. 9, p. 10.

[13] *Bibliographic Formats and Standards*, p. 30.

[14] *MARC 21 Format for Bibliographic Data* (1999 ed.). Washington, DC: Library of Congress. Accessed through the Cataloger's Desktop CD-ROM, Sept. 16, 2003.

[15] *Ibid.*

[16] *Bibliographic Formats and Standards*, p. 4.

[17] *Ibid*, p. FF:3.

[18] *Ibid*, p. 0:3.

[19] OCLC. (Sept. 16, 2003). *OCLC MARC Records: 1993 November-Present*. Online. OCLC. (N.D.) Chap. 5. Retrieved from the World Wide Web:

<http://www.oclc.org/support/documentation/worldcat/records/subscription/default.htm>.

[20] Endeavor Information Systems (Sept. 30, 2003). Endeavor Supports Unicode Standard in Voyager, Encompass. Online. Endeavor Information Systems. (Aug. 4, 2003). Retrieved from the World Wide Web:
<http://www.endinfosys.com/news/unisucc.htm>.

[21] Dorman, D. (2002). OCLC Taps VTLS to Help Migrate WorldCat. *American Libraries*, 2002 (May), 83.

Figures

Figure 1. Web form for adding ETD metadata

Appendices

A. List of significant ETD metadata fields

B. Coding for leader and fixed fields

C. Metadata to MARC map


D. Coding for variable fields

E. Full text of ETD2MARC script

F. Example output record from the Connexion save file

Figure 1: Web form for adding ETD metadata

ETD-Userdata - Microsoft Internet Explorer provided by Texas A&M University Libraries
Address: http://thesis.tamu.edu/

TEXAS A&M UNIVERSITY
ETD SUBMITTAL PROCESS  **THESIS OFFICE**

Welcome to the ETD Submittal Process at Texas A&M University
This site will take you through the steps of submitting your manuscript

Personal Data

Projected Graduation: Semester: Year:

First Name:

Middle Name:

Last Name:

Birthdate: YYYY-MM-DD

Email Address:

Publish my email address: Yes
 No

Degree Data

Defense Date: YYYY-MM-DD

Degree:

Type of Document:

Major Subject:

Department:

Manuscript Data

Capitalize only the first word and proper words in the title

Title:

Appendix A: List of significant metadata fields

ETD-db Variable name	Explanation of variable name
urn	Universal Resource Name
text	Document type (Thesis, Dissertation, etc.)
bdate	Birthdate
title	Title of document
first_name	First name of author
middle_name	Middle name of author
last_name	Last name of author
majorsubject	Major subject area
dtype	Degree type (M.S., M.A., Ph.D. etc.)
abstract	Abstract of document
url	URL of ETD
keywords	Keywords to describe the document

Appendix B: Full text of ETD2MARC script

```
#!/usr/local/bin/perl

## Section 1: Perl module statements.
filesize,birthday
use DBI;
use MARC::Record;
use PDF;
use HTML::Parser;
use HTML::Entities;

## Section 2: Set connection to database.
$DB_TYPE      = "mysql";
$DB_NAME      = "da";
$DB_USER      = "username";
$DB_PASS      = "*****";

$dbh=DBI->
>connect("DBI:mysql:database=$DB_NAME;host=thesis.tamu.edu",$DB_USER,$DB
B_PASS) || die $DBI::errstr;

$sth=$dbh->prepare("SELECT * FROM search where avail =
'Unrestricted'");

$sth->execute or
die "Unable to execute query: $dbh->errstr\n";

my $emailurn;
my $eid=0;
while(@record=$sth->fetchrow_array) {

## Section 3: Assign and format variables.
$emailurn[$eid]=$record[0];
$eid++;
$urn=$record[0];
$author=$record[2];
$degree=decode_entities($record[10]);
$dtype=$record[7];
$title=$record[4];
$doctype = $record[3];
@doctype = split(/ /,$doctype);
$text = $doctype[1];
$rbdate=$record[12];
@sbdate = split(/-/, $rbdate);
$bdate = $sbdate[2];
@author1 = split(/,/, $author);
$authorf = $author1[1]." ".$author1[0];
$abstract=decode_entities($record[13]);
$majorsubject=decode_entities($record[9]);
$year = $record[1];
@ayear = split(/ /,$year);
$year = $ayear[1];
$url = "http://etd.tamu.edu/collections/document.php?daid=".$urn;
$abstract =~s/\x92/'/g; # replace ascii code 146 with normal
apostrophes
$abstract =~ s(<[^>]*>())g; # remove html tags

$keywords = decode_entities($record[5]);
```

Appendix B: Full text of ETD2MARC script

```
chop($keywords);
@akey = split(/, /,$keywords);

## Section 4: Create MARC records.
my $record = MARC::Record->new();
$record->leader('      nam 2200265Ka 4500');
$record->add_fields(
  [ '006', 'm f d ' ],
  [ '007', 'cr n ' ],
  [ '008', ' s'. $year.' xx sb 000 0 eng c' ],
  [ '035', '', '', a=>"(TxCM)$urn" ],
  [ '040', '', '', a=>"TXA", c=>"TXA" ],
  [ '099', '', '', a=>"$year", a=>"$text" ],
  [ '100', '1', '', a=>"$author,", d=>"$bdate-" ],
  [ '245', '1', '0', a=>"$title \/", h=>"[electronic resource]"
, c=>"by$authorf." ],
  [ '260', '', '', c=>"$year." ],
  [ '500', '', '', a=>"Major Subject: '$majorsubject.'" ],
  [ '500', '', '', a=>"Title from author supplied metadata." ],
  [ '502', '', '', a=>"Thesis ($dtype)\-\-Texas A&M University, $year."
],
  [ '504', '', '', a=>'Includes bibliographical references.' ],
  [ '516', '', '', a=>"Text \($text)\." ],
  [ '520', '3', '', a=>"$abstract" ],
  [ '538', '', '', a=>'Mode of access: World Wide Web.' ],
  [ '650', '', '4', a=>"Major $majorsubject." ],
);

foreach $v (@akey){
$record ->add_fields([ '653', '', '', a=>"$v" ]);}
$record ->add_fields([ '856', '4', '0', u=>"$url" ]);

## Save records and FTP them to the Library's system.
$filename = "marc/$urn.dat";
open(OUTPUT, ">marc/$urn.dat");
print OUTPUT $record->as_usmarc();
close(OUTPUT);

use Net::FTP;
$ftp = Net::FTP->new("199.999.999.999", Debug => 0);
$ftp->login("username", '*****');
$ftp->cwd("/pub");
#$ftp->get("that.file");
$ftp->binary();
print "\nTransferring $filename DAT file ... ";
$ftp->put($filename);
print "Done.\n";
$ftp->quit;

}

## Section 6: Close the database call.
$sth->finish;
$dbh->disconnect;
```

Appendix C: Coding for Leader and Fixed Fields

Leader	
Position	Code
00-04: Logical record length	
05: Record status	n: New
06: Type of record	a: Language material
07: Bib level	m: Monograph/item
08: Type of control	_ : No specific type of control
09: Character coding scheme	_ : MARC - 8
10: Indicator count	
11: Subfield code count	
12-16: Base address of data	
17: Encoding level	K: Less than full level
18: Cataloging form	a: AACR2
19: Linked record requirement	
20: Length of the length-of-field portion	
21: Length of the starting-character-position portion	
22: Length of the implementation-defined portion	
23: Undefined	

006 Field	
Position	Code
00: Form of material	m: Computer file
01-04: Undefined	
05: Target audience	f: Specialized
06-08: Undefined	
09: Type of computer file	d: Document
10: Undefined	
11: Govt. publication	_ : Not a government publication
12-17: Undefined	

007 Field	
Position	Code
00: Category of material	c: Computer file
01: Specific material designation	r: Remote
02: Undefined	
03: Color	
04: Dimension	n: Not applicable
05: Sound on medium	
06-08: Image bit depth	
09: File format	
10: Quality assurance targets	
11: Antecedent/source	
12: Level of compression	
13: Reformatting quality	

Appendix C: Coding for Leader and Fixed Fields

008 Field	
00-05: Date entered on file	
06: Publication status	s: single known date/probable date
07-10: Date 1	year
11-14: Date 2	
15-17: Place of publication	xx: no place, unknown, or undetermined
18: Illustration 1	
19: Illustration 2	
20: Illustration 3	
21: Illustration 4	
22: Target audience	
23: Form of item	s: electronic resource
24: Contents 1	b: bibliographies
25: Contents 2	
26: Contents 3	
27: Contents 4	
28: Govt. publication	0: not a government publication
29: Conf. Publication	0: not a conference publication
30: Festschrift	0: not a festschrift
31: Index	0: no index
32: Undefined	
33: Literary form	0: not fiction
34: Biography	_ : no biographical material
35-37: Language	eng: english
38: Modified record	
39: Cataloging source	c: cooperative cataloging program

Appendix D: Metadata to MARC map

ETD variable name	Explanation of variable name	MARC tag(s) variables are mapped to:
urn	Universal Resource Name	035
year	Year of graduation	008, 099, 260, 502
text	Document type (Thesis, Dissertation, etc.)	099, 516
bdate	Birth date	100
title	Title of document	245
first_name	First name of author	100, 245
middle_name	Middle name of author	100, 245
last_name	Last name of author	100, 245
majorsubject	Major subject area	500, 650
dtype	Degree type (M.S., M.A., Ph.D. etc.)	502
abstract	Abstract of document	520
url	URL of ETD	856
keywords	Keywords to describe the document	653

Appendix E: Coding for variable fields

Tag	Ind 1	Ind 2	Field data
035			a (TxCM)urn
040			a TXA c TXA
099			a year a text
100	1		a last_name, first_name middle_name , d bdate-
245	1	0	a title h [electronic resource] / c by first_name middle_name last_name .
260			c year .
500			a "Major subject: major_subject ".
500			a Title from author supplied metadata.
502			a Thesis (dtype)--Texas A&M University, year .
504			a Includes bibliographical references.
516			a Text (text)
520	3		a abstract
538			a Mode of access: World Wide Web.
650		4	a Major major_subject .
653			a keyword
856	4	0	u url

plain text: literally transcribed

bold text: from ETD variable

Appendix F: Example output record from the Connexion save file

Books	Rec Stat	n	Entered	20030916	Replaced	20030916			
Type	a	ELvl	K	Src	c	Audn		Ctrl	
BLvl	m	Form	s	Conf	0	Bio		MRec	
		Cont	b	GPub		LitF	0	Indx	0
Desc	a	Ills		Fest	0	DtSt	s	Dates	2003,
007									c r e n
040									TXA c TXA
035									(TxCM)etd-tamu-2003A-2003032522
099									2003 a Dissertation
100	1_								Won, Jong-Seob, d 1966-
245	10								Intelligent energy management agent for a parallel hybrid vehicle / h [electronic resource] c by Jong-Seob Won.
260									c 2003.
500									"Major Subject: Mechanical Engineering"
500									Title from author supplied metadata.
502									Thesis (Ph. D.)--Texas A&M University, 2003.
504									Includes bibliographical references.
516									Text (Dissertation).
520	3_								This dissertation proposes an Intelligent Energy Management Agent (IEMA) for parallel hybrid vehicles. A key concept adopted in the development of an IEMA is based on the premise that driving environment would affect fuel consumption and pollutant emissions, as well as the operating modes of the vehicle and the driver behavior do. IEMA incorporates a driving situation identification component whose role is to assess the driving environment, the driving style of the driver, and the operating mode (and trend) of the vehicle using long and short term statistical features of the drive cycle. This information is subsequently used by the torque distribution and charge sustenance components of IEMA to determine the powersplit strategy, which is shown to lead to improved fuel economy and reduced emissions.
538									Mode of access: World Wide Web.
650	_4								Major Mechanical Engineering.
653									hybrid vehicle
653									energy management
653									driving cycle
653									intelligent energy management agent
653									torque distribution
653									charge sustenance
856	40								u http://etd.tamu.edu/collections/document.php?daid=etd-tamu-2003A-2003032522
006 fields for Computer Files Computer File									
									Audn f File d GPub