Publication Notes

Title: Guidance Documents for Lifecycle Management of ETDs

Editors: Matt Schultz, Nick Krabbenhoef, and Katherine Skinner

Authors: Daniel Alemneh, Bill Donovan, Martin Halbert, Yan Han, Geneva Henry, Patricia Hswe, Gail McMillan, Xiaocan (Lucy) Wang

Publisher: Educopia Institute, 1230 Peachtree Street, Suite 1900, Atlanta, GA 30309.

Copyright: 2014

This publication is covered by the following Creative Commons License:

Attribution-NonCommercial-NoDerivs 4.0

You are free to copy, distribute, and display this work under the following conditions:

- **Attribution** – You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). Specifically, you must state that the work was originally published as the Guidance Documents for Lifecycle Management of ETDs, and you must attribute the copyright holder as the Educopia Institute.

- **Noncommercial** – You may not use this work for commercial purposes.

- **No Derivative Works** – You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you get permission from the copyright holder. Your fair use and other rights are in no way affected by the above.

The above is a human-readable summary of the full license, which is available at the following URL:

[http://creativecommons.org/licenses/by-nc-nd/4.0/](http://creativecommons.org/licenses/by-nc-nd/4.0/)
# Metadata for ETD Lifecycle Management

Daniel Alemneh (University of North Texas)

## Topics Covered
- Roles of metadata in facilitating the ETD lifecycle.
- Methods to capture metadata manually and automatically.
- Examples of programs using metadata to enhance ETD access.
- Strategies to manage metadata over time.

## 6.1 Introduction

Electronic theses and dissertations (ETDs) are an important output of the research cycle. Since the late 1990s, ETDs have played significant roles, not just as new forms of scholarly communication, but also as drivers for the development of institutional repositories and digital libraries in general. The successful management of ETDs requires effort throughout the entire lifecycle to ensure that ETDs are preserved and made accessible in a manner that today’s users expect and that tomorrow’s users will find useful. As described in this guidance document, the creation and maintenance of information about the ETD files (including technical, functional, and descriptive metadata) is a key component of this effort.

Although the term “metadata” is used differently in different communities, metadata is usually defined as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO 2004). This document first identifies ETD metadata practices at different institutions and discusses the critical role of metadata in facilitating ETD lifecycle management. It then outlines some of the most important metadata elements to capture and discusses stakeholder roles and responsibilities in the creation of this information.

Some metadata elements are vital for ETD lifecycle management activities; others might be considered optional. Some metadata fields can be “extracted” from the files using software tools; other fields need to be filled out by hand. This document will recommend implementation strategies based on current ETD metadata best practices and standards.

## 6.2 Metadata Roles in ETD Lifecycle Management

Advancements in digital technologies shape the creation, access, use, and preservation of information resources in profound ways, including for electronic theses and dissertations. Although most ETDs still resemble their print equivalents and are text-based PDF files, supporting technologies have matured to allow for the creation and incorporation of complex and dynamic resources in an ETD.

In view of the central role played by metadata in digital libraries, the following section highlights some trends that impact ETDs’ curation as well as metadata best practices at large:
• Heterogeneous, multimedia ETDs instead of text documents.
• Complex retrieval systems instead of matching queries and document representations.
• Visualization of the information space instead of a ranked list of search results.
• Human information behavior instead of information needs.
• Access restrictions requested by the author, school, or research sponsor instead of applying a permanent status for the entire lifecycle.
• Users as both creators and consumers of information instead of one or the other.
• Long-term preservation of digital assets of unstable digital objects.
• Deployment of new cataloging standards in libraries, including RDA (Resource Description and Access) and the exposure of RDA-based data in the linked data cloud instead of traditional cataloging tools such as MARC (MACHINE-READABLE CATALOGING) and AACR2 (Anglo-American Cataloguing Rules, 2nd revision).

6.2.1 Structural Metadata
Different disciplines have different ETD structures and requirements. Accordingly, institutional repositories typically support a variety of digital formats, as determined by the ETD program and documented in ETD policies. Candidates for a music doctoral degree, for example, may be required to submit a text-based dissertation, usually an analysis of a composition or a particular composer’s works. This text document may also be accompanied by recitals that demonstrate the candidate’s instrumental/vocal virtuosity and a lecture recital that combines a performance and description of the dissertation topic. Depending upon the policies of an institution, these may be recorded and stored as part of the ETD to provide additional evidence of the candidate’s performance and fulfillment of academic candidacy requirements.

Accompanying materials in the ETD landscape are not limited to performances. Visual forms (geologic diagrams or maps, high resolution images of art objects, videos of lab observations, etc.) accompany dissertations in fields as diverse as environmental science, stratigraphic geology, art, and biology. ETD Programs and policies have taken different approaches to these additional objects—some separate the accompanying materials from the text of the ETD and some do not accept them at all. Increasingly, metadata is used as a way to bind these resources together, allowing academic institutions to provide and preserve multiple-format ETDs that include content such as video, audio, and datasets (Beagrie and Pink 2012).

Although institutions often require electronic theses or dissertations for graduation, some accompanying materials remain in their physical formats. Over the next decade, more theses and dissertations will be digitized retrospectively. Ensuring digital access to accompanying materials will be problematic for those accustomed to a single format and/or the digital representation of complete content. By employing appropriate metadata elements to record and link various characteristics and relationships, institutions can integrate ETDs’ associated contents.

As described in the Managing the Lifecycle of ETDs: Curatorial Decisions and Practices, such integrations of complex content objects have implications for how ETDs are curated. A review of the current landscape in digital libraries and emerging trends shows that there is no shortage of opinions on the role
of metadata in digital resources lifecycle management (Day 2006; Lavoie 2004). Various emerging Web applications – driven by semantic web technologies such as the Web ontology language (OWL), the resource description framework (RDF), semantic Web rules language (SWRL), and other members of the World Wide Web Consortium (W3C) family of specifications – offer powerful data organization, combination, and query capabilities.

6.2.2 Supporting Access Restrictions and Embargos
Limiting access to ETDs according to a student’s campus, either geographically or by user identification, is a key intellectual property concern of ETD programs, and it is addressed with metadata. As stated in the Guide to Access Levels and Embargoes of ETDs, different institutions may have different access restriction polices that can be applied to all or part of a work. However, there are different options for handling restricted ETDs. Most institutions usually have a delay (1-5 years) and then the restricted ETDs move to public access. Depending upon university policies and student requests, the ETD and/or its metadata may be restricted.

To facilitate this, the metadata should contain information about why and for how long the ETD is restricted or embargoed. If the metadata is also restricted and/or embargoed, this should be coded in the metadata too. If this information is recorded in a standard way, it can be used by repository software to determine the status of an individual object and to change that status over time: e.g., if an ETD is embargoed for three years, the metadata should include information sufficient to allow a repository system to know the date upon which the embargo is lifted.

6.2.3 Facilitating Intellectual Property Rights Management

As discussed in the Briefing on Copyright and Fair Use Issues in ETDs, because ETDs capture the research efforts of students, higher education institutions have a responsibility to provide the best possible guidance on students’ intellectual property rights. Among many stakeholders in the ETD lifecycle, libraries play an active role in the long-term stewardship of these resources. While students hold the copyright to their ETDs, they generally grant a library the right to preserve and provide ongoing access to their ETDs.

In light of the cross-disciplinary and cross-institution usage of ETD research, expressing rights and policy statements via metadata is vital for facilitating distribution of ETDs via institutional repositories. However, as stated in Briefing on Copyright and Fair Use Issues in ETDs, issues of copyright and intellectual ownership have been identified as serious concerns for most universities. Authors normally hold the intellectual property rights in the content of their ETDs. It is important to document the minimum core rights information that a repository must know, and what rights or permissions it has in order to carry out actions related to objects within the repository. These may be generally granted by copyright law, by statute, or by a license agreement with the rights holder. US copyright law, for example, governs all items published since 1923, which may differ from other countries. Knowing the birth and death dates of the creator and the year in which the ETD was created will help to calculate and determine the copyright status. For example, if the creator designated his/her estate as the copyright
owner upon his/her death, performance and use of the compositions is the right of the composer and his/her estate until seventy-five years after the composer’s death.

Moreover, in light of transitional events in ETDs’ lifecycle (embargo releases, redactions, and other possible preservation actions), rights metadata information needs to track and document the changes continually. New rights information may be provided or discovered by the copyright owner, users, or other parties. Also, new legislation and policies at national and/or institutional levels will likely require changes in rights metadata information in order to reflect the most current right status of the ETDs.

### 6.2.4 Facilitating Preservation Activities

Recognizing the critical role of metadata in any successful digital lifecycle management strategy, institutions that take responsibility for digital objects should also implement a metadata-based approach to ensuring long-term access. In addition to the descriptive metadata, (which describes the intellectual entity and supports discovery and delivery of ETD content), preservation metadata provides provenance information, documents preservation action, identifies technical information, and helps in verifying the authenticity of digital objects.

Essentially, the ETD metadata needed to support the full range of digital preservation activities can be loosely categorized into the following possibly overlapping groups:

- **Provenance metadata**: records the origin or provides an historic context or source provenance, such as specifying the analog source material for a digital derivative.
- **Structural metadata**: captures physical structural relationships, such as which image is embedded within which file, as well as logical structural relationships, such as page order, in born digital or digitized ETDs.
- **Technical metadata**: captures format-specific technical information that applies to any file type, including information about the software and hardware on which the digital object can be rendered or executed, as well as checksums and digital signatures to ensure fixity and authenticity.
- **Administrative metadata**: provides provenance information regarding who has cared for the digital object and what preservation actions have been performed on it. It also provides rights and permission information that specifies embargoes and access of ETDs and which preservation actions are permissible. As discussed in *Briefing on Copyright and Fair Use Issues in ETDs*, rights take different forms and some rights are covered by laws, like copyright law, which can be different in different jurisdictions. There are also contractual rights (e.g. licenses) that are governed by contracts (terms and conditions) between parties.

Although there is no “one-size-fits-all” solution to the digital preservation problem, the role of metadata in ensuring long-term access and management has been analyzed by many researchers. The PREMIS (PREservation Metadata Implementation Strategies) working group holds Preservation Health Check workshops and implementation fairs on a regular basis with real life examples of preservation metadata. Such a forum can help early adopters, practitioners, and other stakeholders to come together and refine
theoretical assumptions in order to build a better shared understanding of what, why, and how preservation metadata are collected and created (PREMIS 2014).

### 6.2.5 Describing ETDs
Because dissertations must constitute original research, each is unique to the bibliographic world. As such, each dissertation receives original cataloging and metadata descriptions (Alemneh and Hartsock 2014). In many cases users want to know what has been done in a particular discipline or topic area (Voorbij 1998). Although consulting the bibliographic detail or abstract alone is inadequate for scholarship, the complete metadata description can convey sufficient information about ETDs to give users confidence that the ETD will be relevant to their research areas.

Metadata elements describe characteristics or aspects of an object or digital resource. With enhanced metadata-based and subject-specific search mechanisms, it is now easier than ever to access, use, and reuse scholarly works and associated data that have not been available through traditional publishing methods.

Assigning appropriate metadata to ETDs can improve discoverability by increasing their visibility. To describe digital resources accurately, metadata creators try to follow, as closely as possible, the thinking of the creator/author and also to anticipate what users might want to discover and how they will retrieve the information. As noted by many researchers, one of the key issues for information retrieval and all other content-based text management applications is document indexing (Cleveland and Cleveland 2013). The generation of accurate indexing terms is fundamental to the discovery, use, and reuse of digital resources. An index term is simply a systematic representation of an information-bearing object (text, image, audio, video, etc.) that points users to specific items on topics of interest. In other words, it is an information retrieval tool.

Indexing enhances the accessibility and value of a resource, provided that it is based on a thorough analysis of that resource. Similarly, subject metadata conveys what an ETD is about. In addition to the needs and abilities of the searchers, the subject terms needed to sufficiently create access to ETDs depend on the content that is indexed. A good index helps users find what they need, even when they are not sure of what they are looking for. Lancaster (2003) noted that in the rapidly growing information environment, unidentified and disorganized content, however useful it might be, is at risk of being rendered undiscoverable, and thus obsolete. Figure 6-1 depicts the most common forms of document representation that are familiar from the classic model of information retrieval.

Figure 6-1. The classic Information Retrieval model (modified from Bates 1989)
Many information retrieval researchers agree that in light of the continual evolution of users’ information seeking behavior systems should be sufficiently flexible to allow users to adapt to the current environment (Bates 1998; Ellis 2004; and Kuhlthau 2006). Search engines generate results by matching search terms entered with terms in the content, usually referred to as full-text searching. “Among young academic researchers in Sweden, Google was the most-used starting point for searching scientific information” (Haglund and Olsson 2010). This form of searching has shortcomings, mainly in precision. Precision refers to the proportion of the relevant documents retrieved in a search to the total number of documents retrieved in a search. Increasingly, search engines use metadata to enhance precision and improve their search results. Similarly, many researchers noted that in comparison to ten years ago the use of metadata by Google has increased considerably (Beall 2010).

In contrast, many information scientists still argue that we have not developed comparably mature tools for exploratory search – information seeking where users do not have a known target document and may not even have a well-established information need. “Only in the last few years have we seen an emerging program of human–computer information retrieval (HCIR) that brings interactive techniques to bear on more sophisticated information seeking tasks” (Ardo 2011).

### 6.3 Capturing Metadata

Assigning appropriate metadata to ETDs can improve discoverability by increasing their visibility (Ivanovic et al. 2012). There are data elements that serve a function both for the users’ tasks of search and retrieval (bibliographic records) and for digital preservation purposes (preservation metadata). Users are more interested in the content and the subjects, than in what format the objects are delivered. Many commentators agree that the most useful metadata about a digital object are the subjects or keywords, since they explicitly describe what the object is about (Schwing et al. 2012). A number of researchers (Peterson 2010; Spiteri 2011) have analyzed content indexing (especially subject indexing) and described the general information seeking behavior of users. Many agreed that the two major reasons why users experience problems with subject access are the quality and application of subject indexing on the one hand, and the complexity of users’ knowledge and information literacy skills required for successful subject access on the other. To maintain the consistency of search results and high recall of available resources, it is critical to ensure the quality of the keywords and taxonomies used to index heterogeneous digital resources within digital libraries.

#### 6.3.1 Metadata by Librarians

Catalogers and information professionals have always done some assessment of material. To fully understand what a good index is, it is necessary to be both micro- and macro-minded. On the micro level, we concern ourselves with the specific mechanics of creating an index term. On the macro level, indexing is part of the larger context of an information retrieval system. Retrieval of information involves the user expressing an information retrieval request that incorporates terms from the common vocabulary to match stored records.

In this regard, controlled terms provide a broad navigational tool for browsing through digital content and digital library collections. Controlled vocabulary terms ensure indexing consistency and enhance retrieval precision across all digital resources. In many institutions, catalogers adapt existing workflows
and procedures to handle ETDs. The creation of digital cataloging and metadata workflows is a good opportunity to implement general cataloging policy changes.

One response is outsourcing some segment of metadata/bibliographic processing or related technical service operations. This strategy is increasingly seen as a way to cut technical service costs and enhance efficiency. However, outsourcing requires investment in quality control, potentially taking the form of experienced or professional librarians to monitor output.

Another response is opening new positions that facilitate metadata creation: scholarly communication librarian, repository librarian, digital curator, copyright librarian, semantic technologies librarian, etc. Regardless of their labels or names, the primary and common responsibilities of such positions include formulating plans for moving their libraries forward to meet the challenges of changing modes of scholarly communication. This usually includes an active role in promoting open access and developing a metadata strategy for the library and communicating the implications of such strategies to the university community. Their liaison with faculty and students provides opportunities to integrate digital contents and digital repositories into the learning and research mission of their institutions.

6.3.2 Author-Supplied Metadata
In addition to metadata created by professionals, incorporating author-supplied keywords can enhance descriptive metadata. Some libraries (including early ETD adopters such as Virginia Tech) incorporate only author-supplied keywords. However, considering the growing interdisciplinary nature of higher education and the importance of topical approaches for ETD users, the subject matter may not be sufficiently captured by authors’ supplied terms alone.

In view of the constant changes in users’ requirements, access to ETDs relies on a seamless discovery process that offers multiple options to users. Incorporating both controlled and natural approaches to the subject matter in an ETD collection, results in high-level descriptions and representations. Alemneh and Rorissa 2012 noted that such approaches (including using author supplied terms and social tags as possible sources of metadata) add value and enhance user’s ability to find, access, use, and re-use digital objects.

6.3.3 Automatic Extraction
There are a growing number of metadata extraction tools that enable works to be automated either through batch processing or processing on an individual basis as required. Automated metadata extraction is particularly useful for collections like ETDs. For example, the National Library of New Zealand’s Metadata Extraction Tool can programmatically extract preservation metadata from a range of file formats including PDF documents, image files, sound files, Microsoft Office documents, and many others. Although most such tools are designed to facilitate self-documenting and preservation activities over time, they can also be used for other tasks such as the extraction of metadata for resource discovery.

In addition to technical metadata, there are many other vital ETD characteristics that need description, tracking, and integration. Researcher profiles (or author/contributor metadata) can often be collected from an institution’s human resources system (e.g., Banner, PeopleSoft). In most institutions, these
systems are tied to the author’s plan of study and can supply additional information such as academic
discipline, committee members, and document type.

There is growing consensus among academic and research communities about the critical need to link
publications and underlying data. In this regard, the current customized version of DSpace-CRIS (Current
Research Information System) can be cited as one of the emerging new add-ons that enable the
ingestion, storage, display, and management of metadata and full text for ETDs and other research
entities. Using such tools, items from native system and/or new resources from different systems, can
be linked to each other using auto-complete and auto-lookup functions in the submission edit phase.

Considering the emerging multi-format/part structures of ETDs, tools that simplify a smooth integration
between local resources and other entities are particularly useful. Quality metadata plays a significant
role in integrating and contextualizing all of these heterogeneous entities. In turn, this adds greater
value to each individual component and facilitates visibility, discovery, and understanding of the overall
research agenda.

6.3.4 User-Generated Terms – Post-Ingest Activities

Although different metadata schemas are often complementary, good keyword terms help users find
what they need, even when they are not aware of their needs. In the quest for better discoverability of
existing digital resources, libraries implement systems to enrich traditional catalog and metadata
records with additional user-supplied terms and descriptions.

A growing number of institutions are assessing the potential for user-supplied tags or folksonomy to
complement established controlled vocabularies in a diverse and collaborative environment. A
gfolksonomy is any system that allows users to tag their favorite digital resources using natural language
words. In a socially constructed metadata paradigm, users not only search/browse, access, and use
content but also proactively participate in its production and description by tagging, rating, reviewing,
highlighting, and recommending (Alemneh and Rorissa 2014; Smith 2008).

With the increasing popularity of social tagging systems, Sue Yeon Syn and Michael B. Spring (2013)
explored and analyzed social tagging systems as a mechanism to allow nonprofessional catalogers to
participate in metadata generation. The results suggest that user tags successfully identify the terms
that represent the topic categories of web resource content. However, user-generated terms have been
suggested as a lightweight way of enhancing descriptions of digital contents and improving their access
through broader indexing. Trant (2008) summarized both the negatives and positives of folksonomies.
On one hand, critics point to the fact that it is an uncontrolled vocabulary and leads to less effective
information retrieval. On the other hand, proponents of the concept of crowd sourcing point to the fact
that it is user-friendly and enables personalized information retrieval by users.

As folksonomies are in a continual state of flux, they are better able to accommodate current
terminology and concepts than traditional indexing tools and systems such as the Dewey Decimal
Classification and the Library of Congress Subject Headings. Traditional approaches share a basic
problem: the potential users of information are disconnected from the process used to describe that
information.
Combining traditional indexing systems with folksonomies is the solution for delivering a richer digital library user experience. However, socially constructed metadata approaches have not been effectively
implemented so as to augment legacy library metadata. The problem could partially be attributed to the absence of a conceptual metadata framework that could serve as a theoretical basis for a better understanding of the possible uses of Web 2.0 services in libraries (Lagoze 2010). There are small but growing initiatives in various libraries that have effectively engaged the user community from transcribing content to correcting OCR output. Tapping user knowledge does not mean that the information professionals curating a collection in a production environment would be able to step back from the collection after the core metadata was added. As described by many commentators and depicted in Figure 6-2, involving all possible stakeholders and building a dedicated group of contributors requires an investment of repository staff time, as well as strategies for promoting data ownership, ensuring quality, and producing seamless integration.

6.4 Best Practices for ETD Metadata
Maintaining usable ETDs requires high-quality metadata about those digital objects. An effective metadata management approach improves consistency, clarity of data lineage, and relationships between and among objects so that institutions can better integrate resources. There are many metadata practices that are implemented by the ETD community at national, regional, and even international levels. A definition of what constitutes minimal, good, and optimal metadata – regardless of format or schema – depends on many factors. These factors can differ from country to country and from institution to institution.

Quality metadata plays a significant role in facilitating the establishment of a union catalog, which is an indispensable element of library networking and resource sharing. Some countries coordinate activities related to ETDs and work with institutions of higher education to ensure high quality metadata. While ETDs are maintained by the institutions at which they were produced, it is possible to give searchers the appearance of a single collection by gathering all the metadata into a central search engine. When a potentially relevant document is found, the systems redirect the user to the institution that houses the document.

A number of institutions and consortia have developed or instituted local metadata best practices. The following sections provide further description. Although the adopting institutions had already validated the practices as “best” by their standards, they may not be suitable for every institution.

6.4.1 International Initiatives
6.4.1.1 Networked Digital Library of Theses and Dissertations (NDLTD)
The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of ETDs. Since its inception in 1996, NDLTD has worked to improve graduate education, increase the availability of student research, empower students and universities, advance digital library technology, and lower the costs of submitting and handling ETDs.

NDLTD has developed a metadata standard especially suited to ETDs (see Table 6-1). NDLTD provided the following metadata elements as a guideline to develop a faithful crosswalk between local metadata standards and a single standard used for sharing information about ETDs.
<table>
<thead>
<tr>
<th>No.</th>
<th>ETD Metadata Elements</th>
<th>MARC Field &amp; Subfield</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>title</td>
<td>245a</td>
<td>Full title for the work as it appears on the title page.</td>
</tr>
<tr>
<td></td>
<td>Alternative title</td>
<td>246</td>
<td>For title, choose the appropriate title qualifier, (main, alternative, added, translated, etc.), preferably from the controlled list. (UNT Libraries, “Title Role Authorities”)</td>
</tr>
<tr>
<td></td>
<td>(Title qualifier)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>creator</td>
<td>100a</td>
<td>Name of Author</td>
</tr>
<tr>
<td>3</td>
<td>contributor</td>
<td>720a</td>
<td>Name of Committee Member, Thesis Advisor, chair</td>
</tr>
<tr>
<td></td>
<td>(Contributor role qualifier)</td>
<td></td>
<td>For Creator / contributor roles, choose the appropriate qualifier, (Committee Member, Thesis advisor, Examiner, etc.), preferably from the controlled list.</td>
</tr>
<tr>
<td>4</td>
<td>subject / keywords</td>
<td>653a</td>
<td>Subjects and/or keywords describe what the ETD is 'about' and enter as many terms as necessary to capture subject content.</td>
</tr>
<tr>
<td>5</td>
<td>description</td>
<td>520a</td>
<td>Content and Physical Descriptions</td>
</tr>
<tr>
<td></td>
<td>abstract</td>
<td>520a</td>
<td>Abstract usually supplied by ETD authors)</td>
</tr>
<tr>
<td></td>
<td>note</td>
<td>5xx</td>
<td>Only include notes when applicable. (Example release note.)</td>
</tr>
<tr>
<td>6</td>
<td>publisher</td>
<td>260a+b</td>
<td>Usually the institution’s name</td>
</tr>
<tr>
<td>7</td>
<td>date</td>
<td>260c</td>
<td>(publication date, graduation date, )</td>
</tr>
<tr>
<td>8</td>
<td>type</td>
<td>Leader 6&amp;7</td>
<td>Type of resources</td>
</tr>
<tr>
<td>9</td>
<td>format</td>
<td>856q</td>
<td>Physical Description, (preferably from the controlled list)</td>
</tr>
<tr>
<td>10</td>
<td>identifier</td>
<td>856u</td>
<td>Unique identifiers include URL of an ETD.</td>
</tr>
<tr>
<td>11</td>
<td>language</td>
<td>008</td>
<td>If the ETD is in multiple languages, include each of them, preferably from the controlled list.</td>
</tr>
<tr>
<td>12</td>
<td>coverage</td>
<td>651 or 690</td>
<td>Coverage information (usually place and time period)</td>
</tr>
<tr>
<td>13</td>
<td>rights</td>
<td>540</td>
<td>Information about rights may include:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Access (level of access that will be allowed to users)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*License (if there is a license or rights)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Holder (usually ETD author is the rights holder)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Statement (information about:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Rights held in and over the resource,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Conditions under which the work may be used, distributed, reproduced, etc.,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Information about how the rights conditions may change over time,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Whom to contact regarding the copyright of the work.)</td>
</tr>
<tr>
<td>14</td>
<td>Degree</td>
<td>502a</td>
<td>Degree information usually includes degree name, level, discipline, and name of institution and academic department granting the degree associated with the work.</td>
</tr>
<tr>
<td></td>
<td>Degree name</td>
<td>502a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Degree level</td>
<td>502a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Degree discipline</td>
<td>710b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Degree grantor</td>
<td>502a; 710ab</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Other*</td>
<td>540</td>
<td>Other optional descriptive metadata elements can be used or added as appropriate:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Relation (can be used when related items are online such as a doctoral recital and the ETD connected to it.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Metadata Information (who creates/updates what, when, etc.)</td>
</tr>
</tbody>
</table>

Table 6-1. Modified NDLTD’s descriptive metadata for ETDs with semantic MARC crosswalk.
The NDLTD metadata is intended to be flexible enough to be used in a variety of current and future representations of ETDs. If feasible, as suggested in Table 6-1, compiling local controlled vocabularies or assigning metadata values from controlled lists facilitates consistency. Whenever possible, while accommodating local requirements, using existing or widely adopted controlled vocabularies promotes even greater interoperability.

The NDLTD metadata has been used by institutions and partners around the world to export their public metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). As noted by Ivanovic, Ivanovic, and Surla (2012), visibility of ETDs can be increased by putting the digital object or its descriptive metadata (or both) into such networked repositories. Although the NDLTD union archive now contains more than two million records of ETDs, based on the total number of ETDs produced around the world (estimated to be close to 1 million per year), it is far from being a comprehensive union catalog for global ETDs.

6.4.1.2 PREservation Metadata Implementation Strategies (PREMIS)
Addressing the preservation and long-term access issues for digital resources is a significant challenge for repositories. A number of researchers noted that the problem of ensuring long-term access to digital information sources is compounded by the fact that most of the sources are not properly organized.

Most agree that extensive metadata is the best approach to minimize the risk of digital objects becoming inaccessible (Alemneh and Hastings 2010).

Accordingly, a number of national and international projects and initiatives attempted to assess the potential role of metadata in preservation management activities. PREservation Metadata Implementation Strategies (PREMIS) has been influential in providing a "core" set of preservation metadata elements that support the digital preservation process.

PREMIS is a core set of metadata elements (called “semantic units”) recommended for use in all preservation repositories regardless of the type materials archived, the type of institution, and the preservation strategies employed (PREMIS 2014). The PREMIS data model consists of five interrelated entities: intellectual, object, event, agent, and rights with each semantic unit mapped to one of these areas:

- Intellectual entities are conceptual and might be called "bibliographic entities." PREMIS does not actually define any metadata pertaining to intellectual entities as there are plenty of metadata standards to choose from, for example NDLTD Descriptive Metadata for ETDs.
- Objects are what are actually stored and managed in the preservation repository. Although descriptive metadata is out of scope, objects can point to descriptions of intellectual entities or the entity itself, in either direction.
- The event entity aggregates information about actions that affect objects in the repository. Events are extremely important for preservation activities, as it is vital to track actions performed on digital objects (such as capture, compression, validation, replication, migration, etc.) for preserving into the future.
• Agents are actors (people, organizations or software) that have roles in events and in rights statements.
• The rights entity aggregates information about rights and permissions that are directly relevant to preserving objects in the repository (PREMIS 2014).

As can be seen in the diagram of PREMIS data model relationships (see Figure 6-3), “objects” can have relationships to other entities in the data model or to each other. However, “agents” can only act on objects through events or rights, not directly. Table 6-2 lists the elements that are properties of the given entity. Most of PREMIS is actually devoted to describing digital objects that include technical metadata. In addition to format and size, it also includes things like fixity, commonly known as “checksums,” and any passwords or encryption that limits the use of an object. Environment details the software and hardware needed to render or use the object. Relationships are classified as structural, how objects comprise the original ETD relate to one another, or derivative, how service or access copies relate to the original ETD.

Although all of the five entities are interrelated, they can be used and implemented independently from each other. Institutions may adopt one or more PREMIS entities. Decisions on how to implement the recommended entities remains entirely up to the individual repository. This allows a repository to implement localized ETD workflows and submission models. Many academic institutions have implemented object, event, and agent entities. For example, the University of North Texas Libraries has built a tool for capturing and providing user access to PREMIS events that are important in a digital objects lifecycle. Such a tool is fairly focused on a part of the PREMIS model (event and agent) and solves a specific problem, which is to capture events in an objects lifecycle (ingest, fixity check, virus check, replication, migration).

Figure 6-3. The PREMIS data model – Version 2.2 (PREMIS 2014)
The rights entity is of particular interest to the ETDs community. Access rights may be generally granted by copyright law, by statute, or by a license agreement with the rights holder. Repository rights only concern rights to preserve, and an institutional repository needs to know what permissions it has to carry out appropriate preservation actions related to ETDs within the repository. For the purpose of the PREMIS Data Dictionary, statements of rights and permissions are encompassed by the rights entity:

- Rights are entitlements allowed to agents by copyright or other intellectual property law.
- Permissions are powers or privileges granted by agreement between a rights holder and another party or parties.

A number of early PREMIS adopters realized that the rights entity lacked the robustness required by various types of digital objects including ETDs. To address such limitations, the PREMIS Editorial Committee has been working on changes and enhancements to the rights entity semantic units. The PREMIS Editorial Committee incorporated several requested changes to the rights entity and published the greatly expanded version of the PREMIS Data Dictionary for Preservation Metadata version 2.2 in May 2012. Based on several other requests from early implementers, the PREMIS Editorial Committee is also working on a major revision for version 3.0 to include rights pertaining to a license, copyright and statutory rights. It is expected that the majority of this information will be captured from the object itself or from the repository system being used and that a minimal amount will need to be provided by hand (PREMIS 2014).

### 6.4.1.3 ProQuest Theses and Dissertations

ProQuest Theses and Dissertations (PQDT), is a database of dissertations and theses, published electronically or in print. PQDT includes nearly 3 million searchable citations to dissertation and theses from around the world from 1743 to the present day. More than 80,000 new full-text dissertations and theses are added to the database each year through dissertations publishing partnerships with 700

---

**Table 6-2. Summary of types of information included in PREMIS by entity type (PREMIS 2014)**

<table>
<thead>
<tr>
<th>Object</th>
<th>Event</th>
<th>Agent</th>
<th>Rights</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Information about the stored digital objects)</td>
<td>(Information about actions)</td>
<td>(Actors that have roles)</td>
<td>(Rights information)</td>
</tr>
<tr>
<td>Object ID (type and value)</td>
<td>Event ID</td>
<td>Agent ID</td>
<td>Rights statement (ID: type &amp; value)</td>
</tr>
<tr>
<td>Preservation level</td>
<td>Event type</td>
<td>Agent name</td>
<td>Granting Agent</td>
</tr>
<tr>
<td>Object characteristics (format, size, fixity, etc.)</td>
<td>Event date/time</td>
<td>Designation (person, software, organization)</td>
<td>Permission granted</td>
</tr>
<tr>
<td>Storage</td>
<td>Event outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environment (Hardware, software)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital signatures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationships</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linking identifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
academic institutions worldwide. Although most ProQuest services are available for purchase, ProQuest has a DTD (Document Type Definition) that describes the XML (Extensible Markup Language) feed delivered to universities for free (ProQuest 2014).

### 6.4.1.4 Statewide Initiatives

A number of regional associations in the US such as the Ohio Electronic Thesis and Dissertation Association (OETDA), the Texas ETD Association (TxETDA), and the Florida ETD Association (FLETDA) seek to increase ETD knowledge of their members through the sharing of state-wide, national, and international best practices. Likewise, regional ETD consortia such as Texas Digital Library (TDL), OhioLINK ETD Center, and State of Florida ETD System, provide resources and services to their member institutions. These associations and organizations also provide networking and learning opportunities, establish standards, promote ETD open access, and more.

The Texas Digital Library (TDL), for example, is a consortium of higher education institutions in Texas that provides shared services in support of research and teaching. To facilitate ETD workflows, the TDL created the Thesis and Dissertation Submittal System (TDSS), more commonly known as Vireo. The Metadata Working Group of the TDL has also developed a descriptive application profile for ETDs in the Metadata Object Description Schema (MODS). Though other metadata formats were discussed and considered, MODS was chosen primarily because it was based on MARC, and it could capture descriptive information (which is analogous to the traditional catalog record) more easily than other schemas, such as Dublin Core. The only extension needed was degree information. With MODS as the descriptive metadata for ETDs, the working group created a TDL ETD MODS application profile. The schema defines 16 top-level elements that are included in the Vireo ETD Submission System and adopted by most TDL members’ institutions (see Appendix A: Texas Digital Libraries (TDL) ETD Metadata Example).

### 6.4.1.5 Non-US ETD Initiatives

There are several ETDs projects outside of the United States. One of the objectives of an ETD program is to provide easy global access to the results presented in ETDs, irrespective of the languages or where the ETD was written. In this regard more and more international professional societies and organizations, such as ASIS&T (Association for Information Science & Technology), are expanding their dissertation awards for dissertations in other languages, explicitly stating that dissertations are welcome in any language. Nevertheless, implementing widely adopted metadata using English in addition to the original languages (e.g. translating titles and abstracts) facilitates and enhances international access.

Among other successful ETD best practice encouragers, the European ETDs collaborative projects can be mentioned as one of the many international efforts that identified metadata for use when describing the content of an ETD repository. The Dublin Core-based recommendation was the result of collaboration between UK institutions that have been involved in developmental work associated with ETDs. There are a number of other initiatives that aim to develop national and regional resource discovery services and search tools to promote the visibility of ETDs at individual member sites. The following four can be cited as success stories, in terms of promoting ETD-specific best practices:
• Cybertesis: A collaborative program (among 32 universities of Europe, Africa and Latin America) which is a portal developed jointly by the University of Chile, the Universites de Lyon, UnMontreal, and Alexandrie, and the University of Geneva for accessing full text ETDs from many countries, including Bolivia, Brazil, Canada, Chile, France, Hong Kong, Mexico, Peru, Spain, and the United States (Global Open Access Portal 2014).

• Database of African Theses and Dissertations (DATAD): The DATAD database contains citations and abstracts for theses and dissertations completed in African universities. Although African theses and dissertations contain local empirical data that is not available in international literature, African research results are rarely indexed in major international databases. In 2000, the Association of African Universities (AAU) initiated and supported efforts toward putting Africa’s research output onto the mainstream of world knowledge via DATAD (AAU 2014).

• UK’s Electronic Theses Online Service (EThOS): Offers a single point of access to all UK theses. EThOS harvests e-theses from institutional repositories and digitizes paper theses free of charge (EThOS 2014).

• Theses Canada: A union catalog of Canadian theses and dissertations, in both electronic and analog formats. Canadian universities participate in the program voluntarily by submitting approved theses and dissertations to Theses Canada (Library and Archives Canada 2014).

Such collaborative national and regional projects support the open access aspirations of many institutions and provide visibility to the work of scholars both within and outside of the countries of origin. Some initiatives (such as EThOS) help institutions to digitize analog copies and return the digitized versions to local institutions for loading onto their respective institutional repositories. More importantly, these organizations coordinate various activities – including the compilation of national and regional union catalogs for ETDs – and enable the participation of many partner institutions, small or large, with or without an institutional repository.

All these collaborative activities are facilitated by means of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which is a protocol for exposure of metadata rather than content. Using the OAI-PMH, individual sites can make their metadata accessible to search providers and discovery services and still maintain complete control over the resources. Supporting the OAI-PMH at individual institutions is the best way to contribute to such endeavors and promote repository interoperability.

Many institutions provide public access to a number of application programming interfaces (APIs) to their ETDs collections that can be used openly by those interested in programmatically accessing data from such systems. A growing number of discovery services, such as Open Access Theses and Dissertations (OATD), work with institutions’ metadata to index only records that are actually ETDs and are freely available online. In addition to harvested metadata, they include a growing amount of full text "snippets" (about the first 30 pages of the thesis and sample images) to help guide searchers.

6.5 ETD Metadata Quality and Management
As demonstrated above, although ETDs are produced at individual institutions, various statewide, national, and international consortia play vital roles in developing best practices for ETDs management.
They also create environments for early ETD adopters to share and promote their best practices with other institutions pursuing ETDs as a catalyst for digital libraries development.

For the last two decades ETDs have played significant roles, not just as new forms of scholarly communication, but as drivers for the development of institutional repositories and digital libraries. In this regard, cultural heritage institutions are making good progress in digital preservation, specifically in the implementation of preservation metadata schema such as PREMIS. ETD-specific lifecycle management guidelines contribute to preserving and ensuring long-term access to digital resources.

Metadata management provides the tools, processes, and environment to enable repositories to answer a number of questions related to the digital object. An effective metadata management approach can

---

**Figure 6-4.** Flowchart for maintaining high quality ETD metadata

---
help institutions improve consistency, clarity of data lineage, and relationships so that they can better integrate resources. Integrating and contextualizing all of the ETDs components along with original research data and pre-and post-publications or performances, adds greater value to each individual component.

6.5.1 Metadata Quality
Maintaining usable ETDs requires high-quality metadata about those digital objects. Metadata quality characteristics depend on various factors, including: metadata record completeness, consistency, accuracy, provenance, conformance to expectations, and other local criteria and known substantive factors. Metadata quality issues are particularly acute if there are multiple institutions participating in collaborative ETD projects where a high level of interoperability is an important element.

Figure 6-4 shows the continuous metadata quality assurance loop. High levels of precision and recall, the two ways in which we measure any information retrieval system, are dependent on many activities that require the involvement of both machine and human interventions. In the figure, arrow line weights reflect the relative importance of each activity. Arrow labels describe physical information retrieval and system administration activities. The central oval represents intellectual activities performed by digital curators. In addition to employing workflows compliant with national and international standards, digital curators must be able to understand how users access ETDs, and then adjust the system for optimum retrieval.

6.6 Summary
The successful management of ETDs requires effort across the entire lifecycle to ensure that ETDs are preserved and made accessible in a manner that today’s users expect and that will be useful to scholars in the future. There are several players over the entire ETDs lifecycle.

Considering the growing interdisciplinary trends in higher education and constant development of information and knowledge management, ETDs demand specialized treatment and characterization that can best capture the semantics and relations of the underlying concepts.

Most academic institutions are interested in making sure that their scholarly output is available to the widest possible global audience. Repositories employ a metadata based dissemination strategy in order to facilitate access, discovery, and use. In order to thoroughly describe ETDs and achieve the required data quality, it is important to engage stakeholders at all stages of the ETD lifecycle. Depending on the roles of the stakeholders, the types of information that should be captured for ETDs and by whom varies, not just at the point of ingestion, but also subsequently, as ETDs often have supplemental files (video, audio, data sets, errata, etc.) and transitional events in their lifecycle (embargo releases, redactions, etc.).

Integrating and contextualizing all ETD components and optimizing the content for search engines adds value and brings greater visibility to individual components across the repository. Most ETD repositories are OAI-PMH compliant, which means that their ETDs can be harvested into the OAI-based union catalog, regardless of the software platform. These multiple approaches enhance an ETD user’s ability to
find available resources while leveraging the benefits of composite applications, mash-ups, and service-oriented architectures.

In addition to the description of required metadata elements and the rationale for why they should be used or adopted, this document also narrates the process of creating the metadata necessary to provide complete access to the users of ETDs. Observations about the ETDs metadata, together with the best practices and their associated framework, are offered in the spirit that they may serve as an implementation roadmap for: creating shareable metadata, ensuring interoperability, and assisting the ETD community in meeting the larger digital curation and lifecycle management challenges. Such practice will have important implications for future researchers, potential ETDs users, and stakeholders engaged in various aspects of digital curation efforts in general.

Here it should be emphasized that the implication of ETD lifecycle management will not be limited to the ETD community or higher learning institutions. Stakeholders benefit from metadata management ranging from the university community to external partners in business, industry, government, and society at large.
Appendix A: Texas Digital Libraries (TDL) ETD Metadata Example

The Texas Digital Library (TDL) provides a digital infrastructure for the scholarly activities of Texas universities. The TDL serves as the center of excellence for the creation, curation, and preservation of digital scholarly information including ETDs. The following list outlines the minimum elements for ETDs descriptive metadata for members of TDL. In addition to the following 16 elements (15 mandatory and 1 optional elements), other valid MODS elements may be included in ETD records as appropriate. Detail information regarding the MODS Application Profile for ETD can be found at TDL home page: http://tdl.org.

Title Information
- title
- subtitle

Name of Author
- type=personal
- namePart=given
- namePart=family
- namePart=date
- roleTerm=Author

Name of Thesis Advisor
- type=personal
- namePart=given
- namePart=family
- namePart=date
- roleTerm=Thesis advisor

Name of Committee Member [Optional]
- type=personal
- namePart=given
- namePart=family
- namePart=date
- roleTerm=Committee Member

Name of Degree Grantor
- type=corporate
- namePart=University Name
- namePart=Department
- roleTerm=Degree grantor

Type of Resource
- typeOfResource=text

Genre
- genre=theses
- authority=marcgt

Origin Information
- dateCreated
- dateIssued

Language
- languageTerm

Physical Description
- form=electronic
- internetMediaType=application/pdf
- digitalOrigin=born digital

Abstract
- authority
- topic
- geographic
- temporal

Subject
- type

Identifier
- type

Location
- url=permalink

Degree Information
- name=Doctor of Philosophy
- level=Doctoral
- discipline=Educational Administration

Record Information
- recordContentSource
- recordCreationDate
- recordChangeDate
- recordIdentifier
Bibliography


Saracevic, Tefko. 2006. “Relevance: a review of the literature and a framework for thinking on the

Schwing, Theda, Sevim McCutcheon and Margaret Beecher Maurer 2012. “Uniqueness Matters: Author-
Supplied Keywords and LCSH in the Library Catalog.” Cataloging & Classification Quarterly 50(8):


Syn, Sue Yeon, and Michael B. Spring. 2013. “Finding subject terms for classificatory metadata from
user-generated social tags.” Journal of the American Society for Information Science and
Technology 64 (5): 964–980.


Concepts, Retrieval, and services. DOI: 10.2200/S00190ED1V01Y200904ICR005.


UNT Libraries. 2014. APIs for UNT Theses and Dissertations.
http://digital.library.unt.edu/explore/collections/UNTETD/api/.

Voorbij, Henk J. 1998. “Title keywords and subject descriptors: a comparison of subject search entries of

Rafael, CA: Morgan and Claypool. doi:10.2200/S00174ED1V01Y200901ICR003.

Information Organization.” First Monday 14 (5).


Zeng, Marcia L., and Lois Mai Chan. 2006. “Metadata Interoperability and Standardization – A Study of
Methodology Part II: Achieving Interoperability at the Record and Repository Levels.” D-Lib
Zhang, Ning, Yuan Zhang and Jie Tang. 2009. “A tag recommendation system for folksonomy.” 
*Proceeding of the 2nd ACM workshop on Social web search and mining*, Hong Kong, China.