

Metadata Schemas/DTDs for ETDs

Introduction

Here we examine the potential metadata elements that we wish to collect when an ETD is submitted online. We look at the default configuration of the qualified Dublin Core register in DSpace, as well as the metadata standards employed by two ETD specific schemas: the ETD-MS (Electronic Theses and Dissertations Metadata Schema) from Virginia Tech and the NDLTD and the TDM DTD (Theses and Dissertations Markup Document Type Definition).

DSpace Default DC Registry Contents

This metadata set comes built into DSpace and can be modified for use in the archive. It is built upon qualified Dublin Core.

Element	Qualifier	Description
contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors.
	advisor	Use primarily for thesis advisor.
	author	
	editor	
	illustrator	
	other	
coverage	spatial	Spatial characteristics of content.
	temporal	Temporal characteristics of content.
creator		Do not use; only for harvested metadata.
date		Use qualified form if possible.
	available	Date or date range item became available to the public.

	accessioned	Date DSpace takes possession of item.
	copyright	Date of copyright.
	created	Date of creation or manufacture of intellectual content if different from date.issued.
	issued	Date of publication or distribution.
	submitted	Recommend for theses/dissertations.
description		Catch-all for any description not defined by qualifiers.
	abstract	Abstract or summary.
	provenance	The history of custody of the item since its creation, including any changes successive custodians made to it.
	sponsorship	Information about sponsoring agencies, individuals, or contractual arrangements for the item.
	statementofresponsibility	To preserve statement of responsibility from MARC records.
	tableofcontents	A table of contents for a given item.
	uri	Uniform Resource Identifier pointing to description of this item.
format		Catch-all for any format information not defined by qualifiers.
	extent	Size or duration.
	medium	Physical medium.
	mimetype	Registered MIME type identifiers.
identifier		Catch-all for unambiguous identifiers not defined by qualified form; use identifier.other for a known identifier common to a local collection instead of unqualified form.
	citation	Human-readable, standard bibliographic citation of non-DSpace format of this item
	govdoc	A government document number
	isbn	International Standard Book Number
	ismn	International Standard Music Number
	issn	International Standard Serial Number
	other	A known identifier type common to a local collection.
	sici	Serial Item and Contribution Identifier
	uri	Uniform Resource Identifier

language		Catch-all for non-ISO forms of the language of the item, accommodating harvested values.
	iso	Current ISO standard for language of intellectual content, including country codes (e.g. "en_US").
publisher		Entity responsible for publication, distribution, or imprint.
relation		Catch-all for references to other related items.
	haspart	References physically or logically contained item.
	hasversion	References later version.
	isbasedon	References source.
	isformatof	References additional physical form.
	ispartof	References physically or logically containing item.
	ispartofseries	Series name and number within that series, if available.
	isreferencedby	Pointed to by referenced resource.
	isreplacedby	References succeeding item.
	isversionof	References earlier version.
	replaces	References preceding item.
	requires	Referenced resource is required to support function, delivery, or coherence of item.
	uri	References Uniform Resource Identifier for related item.
rights		Terms governing use and reproduction.
	uri	References terms governing use and reproduction.
source		Do not use; only for harvested metadata.
	uri	Do not use; only for harvested metadata.
subject		Uncontrolled index term.
	classification	Catch-all for value from local classification system; global classification systems will receive specific qualifier
	ddc	Dewey Decimal Classification Number
	lcc	Library of Congress Classification Number
	lcsh	Library of Congress Subject Headings
	mesh	MEdical Subject Headings

	other	Local controlled vocabulary; global vocabularies will receive specific qualifier.
title		Title statement/title proper.
	alternative	Varying (or substitute) form of title proper appearing in item, e.g. abbreviation or translation
type		Nature or genre of content.

ETD-MS Schema Structure

Data Types

These items define what sort of data can be stored within each element.

Name	Extension	Restriction	Attribute	Enumeration Value	Pattern	Description
freeTextType	string		translated			When a free text field is translated by someone other than the author, that person's name should appear as the value to the translated attribute.
			AG: specialAttrs			
controlledTextType	string		scheme			When the content of a field is controlled, as in subject or date fields, the controlling scheme should be annotated with either or both the name of the scheme and/or a URI describing the controlled object within the context of the scheme.

			resource			
			AG: specialAttrs			
authorityType	string		resource			Each reference to an individual or institution in any field should contain a string representing the name of the individual or institution as it appears in the work. Where possible, the reference should also contain a URI which points to an authoritative record for that individual or institution.
descriptionRoleType	string			note		Optional attributes to qualify the meaning of the description tag. "note" indicates additional information regarding the thesis or dissertation. Example: acceptance note of the department. "release" indicates a description of the version of the work. Should only be used for errata, etc..
				release		
dateType		string			YYYY-MM-DD	As defined in ISO 8601 and the profile recommended for implementing ISO 8601 dates in Dublin Core.

AG = Attribute Group, in this case provided by W3C.

Element Set

These are the actual metadata elements in ETD-MS, along with their datatypes from above and whether they are required or not.

Element	Sub-Element	Data Type	Additional Attributes	Required	Description
title		freeTextType		y	A name given to the resource. In the case of theses and dissertations, this is the title of the work as it appears on the title page or equivalent.
alternativeTitle		freeTextType		n	Alternative title of the thesis or dissertation.
creator		authorityType		y	An entity primarily responsible for making the content of the resource. In the case of theses or dissertations, this field is appropriate for the author(s) of the work. Like other names and institutions, this field should be entered in free text form as it appears on the title page or equivalent, with a link to to an authority record if available.
subject		controlledTextType		y	The topic of the content of the resource. In the case of theses and dissertations, keywords or subjects listed on the title page can be entered as free text. The "scheme" qualifier should be used to indicate a controlled vocabulary.
description		freeTextType		n	An account of the content of the resource. In the case of theses and dissertations, this is the full text of the abstract unless otherwise qualified.
		descriptionRoleType	role		
publisher		authorityType		n	An entity responsible for making the resource available. This is typically the group most directly responsible for digitising and/or archiving the work. The publisher may or may not be exactly the same as thesis.degree.grantor. Like other institutional names,

					this field should be entered in free text form as it appears on the title page or equivalent, with a link to an authority record where available.
contributor		freeTextType		n	An entity responsible for making contributions to the content of the resource. Typical use would be for co-authors of parts of the work as well as advisors or committee members. Co-authors of the entire work would be more appropriate for the creator field.
		string	role		
		anyURI	resource		
date		dateType		y	A date associated with an event in the life cycle of the resource. In the case of theses and dissertations, this should be the date that appears on the title page or equivalent of the work. Should be recorded as defined in ISO 8601 and the profile recommended for implementing ISO 8601 dates in Dublin Core.
type		freeTextType		y	The nature or genre of the content of the resource. This field is used to distinguish the resource from works in other genres and to identify the types of content included in the resource. The string "Electronic Thesis or Dissertation" is recommended as one of the repeatable values for this element. In addition, specify types of content using the standard vocabulary found at: http://dublincore.org/documents/dcmi-type-vocabulary/ . Degree and Education Level are now handled by the thesis.degree field.
format		freeTextType		n	The physical or digital manifestation of the resource. In the case of an electronic thesis or dissertation, this should contain a list of the electronic format(s) in

					which the work is stored and/or delivered. Use the standard MIME type whenever possible (for a list of "registered" MIME types, visit ftp://ftp.isi.edu/in-notes/iana/assignments/media-types/media-types). List as "unknown" if no format information is available, omit if the work is not available in electronic form.
identifier		string		y	An unambiguous reference to the resource within a given context. This can and should be used to provide a URI where the work can be viewed or downloaded. Persistent URNs such as PURLs (http://purl.org/) or Handles (http://handle.net/) are recommended.
language		string		n	A language of the intellectual content of the resource. This should be the primary language the work is recorded in. Portions of the larger work that appear in other languages should use the lang qualifier. See Global Qualifiers. Language names themselves should be recorded using ISO 639-2 (or RFC 1766). If the language is not specified, it is assumed to be English (en).
coverage		controlledTextType		n	The extent or scope of the content of the resource. Should be used for time periods or spatial regions.
rights		freeTextType		n	Information about rights held in and over the resource. Typically, this describes the conditions under which the work may be distributed, reproduced, etc., how these conditions may change over time, and whom to contact regarding the copyright of the work.
degree				n	The degree associated with the work.
	name	freeTextType			Name of the degree associated with the work as it

				appears within the work (example: Masters in Operations Research).
	level	string		Level of education associated with the document. Example: bachelors, masters, doctoral, postdoctoral, other.
	discipline	freeTextType		Area of study of the intellectual content of the document. Usually this will be a department name.
	grantor	authorityType		Institution granting the degree associated with the work. Like other institution names, this field should be entered in free text form as it appears on the title page or equivalent, with a link to an authority record where available.

TDM DTD Specification

This shows the elements and their various sub-elements along with the basic datatype they take and a description taken from the DTD itself. The TDM is designed not only for building the metadata for the item but also laying out the item in full. All references to the layout elements have been removed since we are only interested in storing the true metadata.

Element Path	Content	Description
date	PCDATA	Any date can be entered. For truly precise, the attribute of "notation" allows selection of European/"eur" (day-month-year) or U.S.A./"usa" (month-day-year).
pages	PCDATA	Identification of the page numbers in a citation if desired.
publisher	PCDATA	For identifying the publisher of a work.
editor	PCDATA	Part of a citation, identifies editor(s) of a given work, if desired.

pubPlace	PCDATA	For identifying the place of publication for a work.
volumeissue	PCDATA	Volume and issue of a serial publication.
head		One can specify how to render this tag- e.g., italic.
head.title		For now, limited to appearing only in the head element. See "worktitle."
body		The three main elements of the ETD, the front matter (such as certificate of approval, abstract, titlepages, epigraphs, etc.), primary content (chapters, which are here designated TEI-friendly as "div"), and back (the bibliography and appendix) are included in the "body" element as options (hence the "?" after each) to permit the assembly of the ETD in multiple files. Effectively, there will be one "front.html" and several "div.html" files, and one "back.html" file. So that this same DTD can be used with each, those three main subdivisions are made optional.
body.front		This will be a separate file, logically named as "front.html" in your ETD directory. Each item is separated by a "pb"/pagebreak tag (TEI), and visually in HTML by a horizontal rule/"hr" tag. You may find that several line breaks/"br" are useful to visually separate the material. Use the template file you have been given on disk to assure proper format. Preset, pre-formatted pages are layed out
body.front.titlePage		See Graduate College Thesis Manual, page 3, also examples on pages 15-16. Don't forget to add the "pb"/page break tag at the end. Use docDate for the graduation date.
body.front.titlePage.docTitle	PCDATA	This this the title of the thesis or dissertation. Since some topics may cover foreign, mathematical, author's works, etc., these sub-elements are allowed- and you are encouraged to use them- in order to assure the most precise categorization of your dissertation or thesis.
body.front.titlePage.docTitle.worktitle	PCDATA	Since HTML limits "title" to the "head" element, for now, titles other than "docTitle" (the title of your dissertation or thesis), will use worktitle. The "level" attribute is required, signifies whether it is "m" (monographic- a book, monograph, or other publication under a single autonomous title), "s" (a series

		title), "j" (a journal title), "u" (unpublished title, like a dissertation ;-). Additionally, you can use "type" to specify it is an "abbreviated" version of the title, if it is the "main" title, if it is a "subordinate" title, or, if it is a translation, for instance, it is a "parallel" title.
body.front.titlePage.docTitle.author	PCDATA	identifying an author of a quote work, etc. In other words, an author other than the dissertation/thesis writer, identified a docAuthor. "Name" option is if the given, middle, and surnames are to be specified.
body.front.titlePage.docAuthor	PCDATA	Fill in the dissertation or thesis author according to the sub-elements if desired to be specific as to given, middle, surname found under the "name" element. Otherwise, simple typing of the name with no such delineation is allowed.
body.front.titlePage.docDate	PCDATA	The graduation date (e.g., "May 1999").
body.front.titlePage.thesisadvisor	PCDATA	Identifies the thesis advisor.
body.front.thesiscopyright	PCDATA	The thesis copyright statement is optional, but should be centered and followed by a page break.
body.front.certifapproval	PCDATA	This is the most important part of the front matter in many ways. Candidates are encouraged to scan the actual committee signatures and place them in this section. To enable a more precise layout, the table tag is allowed. Content must conform to the Graduate College Thesis Manual, page 3, and the example pages 17-18. Ending with a pagebreak is advised. Use docDate if graduation date is required.
body.front.dedication	PCDATA	This is the dedication, if any. Graduate College Thesis Manual, page 3.
body.front.epigraph	PCDATA	for front matter, including any epigraph desired according to the Graduate College Thesis Manual, page 4.
body.front.acknowledgements	PCDATA	Anyone you wish to thank would go here, according to the Graduate College Thesis Manual page 4.
body.front.abstract		This should include the full abstract submitted according the the Graduate College Thesis Manual pages 4, 9, and 29. You are encouraged to tag this in detail as many search mechanisms only go as far as an abstract.
body.front.abstract.abstractcover	PCDATA	This is the cover sheet to the abstract. For the doctoral candidate, it should be

		formatted as close to the Graduate Thesis Manual specifications on page 29 as possible (but, of course, you still have to submit a proper print abstract upon deposit). Use docDate for the graduate date.
body.front.abstract.abstracttext	PCDATA	The text of your abstract goes here, see Graduate College Thesis Manual pages 4 and 9. You will also, of course, be handing in a printed abstract for UMI if you are a doctoral candidate. Using the "pre"/preformat tag can guarantee a proper double-spaced printout.
body.front.abstract.abstracttext.worktitle	PCDATA	Since HTML limits "title" to the "head" element, for now, titles other than "docTitle" (the title of your dissertation or thesis), will use worktitle. The "level" attribute is required, signifies whether it is "m" (monographic- a book, monograph, or other publication under a single autonomous title), "s" (a series title), "j" (a journal title), "u" (unpublished title, like a dissertation ;-). Additionally, you can use "type" to specify it is an "abbreviated" version of the title, if it is the "main" title, if it is a "subordinate" title, or, if it is a translation, for instance, it is a "parallel" title.
body.front.abstract.abstractapproval	PCDATA	The line indicating the thesis advisor's approval of your abstract goes here, limited according the Graduate College Thesis Manual page 9. If possible, a scanned image of the signature is desirable. "table" is allowed to enable precise placement of the text and signature.
body.front.toc	PCDATA	Table of contents.
body.front.tablelist	PCDATA	The table list, if applicable, is much like any ordered/unordered list. The preference of the specific department/committee chair is honored. Cosmetic spacing with the horizontal rule/"hr" is permitted. TEI-compliant "pb" is required at the end.
body.front.figurelist	PCDATA	The figure list, if applicable, is much like any ordered/unordered list. The preference of the specific department/committee chair is honored. Cosmetic spacing with the horizontal rule/"hr" is permitted. TEI-compliant "pb" is required at the end.

body.front.symbolabbrevlist	PCDATA	The symbol abbreviation list, if applicable, is much like any ordered/unordered list. The preference of the specific department/committee chair is honored. Cosmetic spacing with the horizontal rule/"hr" is permitted. TEI-compliant "pb" is required at the end. In case there are symbols rendered only as images, the "img." tag is allowed.
body.front.preface	PCDATA	The preface is optional, and should conform with the Graduate College Thesis Manual guidelines for content on page 5.
body.back		This is all the matter following the text proper, bibliography, appendix(es, if any), and the notes.
body.back.notes	PCDATA	These are the end, or foot, notes. The notes follow an unordered list, preferably. If they were to be identified by number, as revisions occur the chance of broken references (a href="#mycitedsource" rel="note") increases. It can also be more intuitive to remember notes by topic for editing and cross-referencing ease. However, the ordered, or numbered, list is also allowed, for those who prefer the conventional methods. In this case, "sup" is also included so a conventional superscript numeral can designate the link to the foot/endnote.
body.back.bibl	PCDATA	The bibliography is entered here, to the degree of detail desired.
body.back.appendix	PCDATA	This model for the appendix accounts for basic text, links, images, and some formatting. If you have more than one appendix, use the "id" attribute to differentiate. "Argument," instead of "hi" is the tag for identifying revisions (n) and any actual argument or rhetorical structures (id).
body.back.appendix.argument	ANY	to which attributes corresponding to the arguments identified above (id) and/or specific revision-related material (n) that occurs in an appendix (e.g., a committee member states that given material should be moved to an appendix, or material is too big for just a note). As TEI does not accept "hi" in the appendix, this tag is used instead.