ETD MS v2.0: A New Schema Draft for Electronic Theses and Dissertations

Lamia Salsabil[1], Jian Wu[1], Edward Fox[2], William A. Ingram[3], Ana Pavani[4]

[1] Computer Science, Old Dominion University

[2] Computer Science, Virginia Polytechnic Institute and State University

[3] University Libraries, Virginia Polytechnic Institute and State University

[4]Electrical Engineering, Pontifícia Universidade Católica do Rio de Janeiro

Abstract

The growth of electronic theses and dissertations (ETDs) in academic repositories requires comprehensive and robust schemas for compliance with the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles. Although Dublin Core and ETD MS v1.1 have been in use for years, they are not sufficient to hold all the document-level metadata and content-level metadata. The latter describes objects that are usually resulted from artificial intelligence-based (AI-based) methods, and are increasingly crucial to facilitate the development of more competitive models to mine scientific knowledge from ETDs, and scholarly publications in general. Organizing and incorporating content-level metadata into the ETD schema will increase the data reusability, reproducibility, and comparability across various studies. This paper addresses this critical need by proposing a new ETD schema draft consisting of: a core schema that enhances the current ETD MS v1.1 schema and an extended schema that describes content derived from ETDs using AI-based methods.

Our work builds upon a comprehensive analysis of two existing metadata standards: Dublin Core (DC) and ETD-MS v1.1. DC represents a universal and widely adopted metadata schema for digital documents in general. However, it lacks elements designed to capture the special fields of ETDs, such as committee chair, degree level, and the author's academic program.

Although ETD-MS v1.1 offers a specific schema for ETDs, the schema lacks separate fields for key information such as discipline and department, making it difficult to categorize ETDs precisely. Additionally, it does not include all necessary metadata fields that could appear in ETDs, such as rights management details, license information, and metadata about embedded objects (figures, tables, etc.). Furthermore, certain fields are ambiguous, such as dc.type. Finally, ETD-MS v1.1 does not include data derived from the content using AI-based methods.

We analyze a corpus containing metadata from 500 ETDs selected from over 500K ETDs that were collected by crawling 114 US university libraries. We aim to identify the strengths and limitations of ETD MS v1.1 in capturing ETD-specific information. The core schema enhances the existing ETD MS v1.1 schema by incorporating essential ETD metadata elements that were previously missing and focuses on capturing fine-grained document-level metadata. The schema defines 6 core entities, with a total of 51 data fields that describe different aspects of an ETD. The core entity, named "ETDs", captures detailed metadata about each thesis or dissertation, including title, author, abstract, institution, and degree information. Other entities that provide additional details relate to the "ETDs" entity through foreign keys. For example, the "Texts" entity stores metadata of the full text.

The extended schema describes fine-grained data extracted from ETDs using AI-based methods such as classification, segmentation, topic modeling, and summarization. It contains 28 entities, including those from the core schema. These entities cover a broader range of information, with a total of 284 data fields. For example, the "Classifiers" entity represents each AI model that predicts the categories for ETDs or objects based on a specific classification system.

This work lays the foundation for storing, managing, and manipulating ETD data for both digital libraries and ETD mining research across all academic domains. Under this schema, we will build a proof-of-concept database containing 1000 ETDs spanning 50 years. We will describe the tools adopted to obtain information to populate the database. In the future, we will revise the schema draft based on feedback from a wide spectrum of potential ETD users and maintainers. We will build portable and easy-to-use software packages that automatically obtain the information specified in the schema from multiple sources and build an extended database under the new schema.