ETD-MS v2.0: A Proposed Extended Standard for Metadata of Electronic Theses and Dissertations

Lamia Salsabil[1], Jian Wu[1], William A. Ingram[2], Edward Fox[2]
[1]Old Dominion University, [2]Virginia Polytechnic Institute and State University, USA

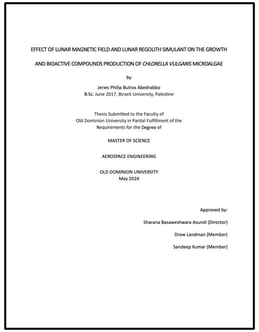
27th International Symposium on Electronic Theses and Dissertations (ETD’24)

ETD 2024
Global Visibility
Livingstone · Zambia

Key Concepts: ETD, ETD Metadata, and Document Schema

What is ETD?

Electronic Theses and Dissertations (ETDs) are electronic versions of theses and dissertations



EFFECT OF LUNAR MAGNETIC FIELD AND LUNAR REGOLITH SIMULANT ON THE GROWTH

AND BIOACTIVE COMPOUNDS PRODUCTION OF *CHLORELLA VULGARIS* MICROALGAE

by

Jeries Philip Butros Abedrabbo
B.Sc. June 2017, Birzeit University, Palestine

Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

AEROSPACE ENGINEERING

OLD DOMINION UNIVERSITY
May 2024

Approved by:

Sharana Basaweshwara Asundi (Director)

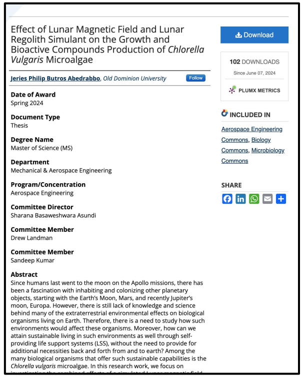
Drew Landman (Member)

Sandeep Kumar (Member)

ETD - https://shorturl.at/utZiP

What is ETD Metadata?

ETD metadata refers to the structured information that describes Electronic Theses and Dissertations (ETDs)



Effect of Lunar Magnetic Field and Lunar Regolith Simulant on the Growth and Bioactive Compounds Production of *Chlorella Vulgaris* Microalgae

Jeries Philip Butros Abedrabbo, *Old Dominion University*

Date of Award
Spring 2024

Document Type
Thesis

Degree Name
Master of Science (MS)

Department
Mechanical & Aerospace Engineering

Program/Concentration
Aerospace Engineering

Committee Director
Sharana Basaweshwara Asundi

Committee Member
Drew Landman

Committee Member
Sandeep Kumar

Abstract
Since humans last went to the moon on the Apollo missions, there has been a fascination with inhabiting and colonizing other planetary objects, starting with the Earth's Moon, Mars, and recently Jupiter's moon, Europa. However, there is still lack of knowledge and science behind many of the extraterrestrial environmental effects on biological organisms living on Earth. Therefore, there is a need to study how such environments would affect these organisms. Moreover, how can we attain sustainable living in such environments as well through self-providing life support systems (LSS), without the need to provide for additional necessities back and forth from and to earth? Among the many biological organisms that offer such sustainable capabilities is the *Chlorella vulgaris* microalgae. In this research work, we focus on

102 DOWNLOADS
Since June 07, 2024

PLUMX METRICS

INCLUDED IN

Aerospace Engineering Commons, Biology Commons, Microbiology Commons

SHARE

ETD Metadata - digitalcommons.odu.edu/mae_etds/378/

What is ETD Metadata Schema?

ETD Metadata Schema is a structured framework that defines the specific metadata elements and their relationships for ETDs

Example: ETD-MS v1.1

Related Work

- Dublin Core (DC)
  - Dublin Core metadata standard first proposed in 1995
  - A widely used metadata standard for describing resources of any type
  - 15 metadata elements
  - Cover fundamental aspects of a resource, such as title, creator, subject, and date
  - Goals [1]
    - Simplicity of creation and maintenance
    - Commonly understood semantics
    - Interoperability
- ETD-MS v1.1 [2]
  - A metadata standard for describing an ETD
  - 22 metadata elements
  - Uses DC metadata elements and new element specifically for theses

[1] Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery (No. rfc2413)
[2] Hickey, T., Pavani A., Suleman, H. ETD-MS v1.1: Metadata standard for electronic theses and dissertations. https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html

Related Work

- TDL Descriptive Metadata Guidelines for ETDs, Version 2.0 [3]
  - A detailed ETD MODS schema, later simplified into DC for broader compatibility
  - Provides basic interoperability
  - Detailed ETD descriptions with fields for Author, Thesis Advisor, Committee Member, Author Identifier (e.g., ORCID), and Embargo details
- Mining ETD Content with Advanced Techniques
  - Rich information in ETD content, including domain knowledge, scientific findings, and technical details
  - Recent Advancements to mine the content of ETDs
    - Multimodal models for ETD page classification [4]
    - YOLO-based methods for detecting figures and tables [5]
    - LDA for identifying core ETD themes [6]

[3] Rushing, A., Koenig, J., Mitchell, A., Moen, W., Strawn, T., &amp; Thomale, J. (2008). Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations, Version 1.0. Prepared for and published by the Texas Digital Library

[4] Choudhury, M. H., Salsabil, L., Ingram, W. A., Fox, E. A., &amp; Wu, J. (2024, March). ETDPC: A Multimodality Framework for Classifying Pages in Electronic Theses and Dissertations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 21, pp. 22878-22884). https://doi.org/10.48550/arXiv.2311.04262

[5] Kahu, S. Y., Ingram, W. A., Fox, E. A., &amp; Wu, J. (2021). Scanbank: A benchmark dataset for figure extraction from scanned electronic theses and dissertations. arXiv preprint arXiv:2106.15320. https://doi.org/10.48550/arXiv.2106.15320

[6] Lamba, M., &amp; Madhusudhan, M. (2019). Mapping of ETDs in ProQuest dissertations and theses (PQDT) global database (2014-2018). Cadernos BAD, 1, 169-182. https://doi.org/10.5281/zenodo.3599788

Why is ETD Metadata Schema Important?

Findability

Effective search and retrieval of ETDs

Accessibility

Accessibility and management of ETDs

Interoperability

Seamless data exchange, cross-repository searches and unified access

Reusability

Reusability and validation of research findings

Comprehensive and robust schema for compliance with the **FAIR** (Findability, Accessibility, Interoperability, and Reusability) principles [7]

[7] da Silva Santos, L. O. B., Burger, K., Kaliyaperumal, R., & Wilkinson, M. D. (2023). FAIR data point: a FAIR-oriented approach for metadata publication. Data Intelligence, 5(1), 163-183. https://doi.org/10.1162/dint_a_00160

Limitations of Existing Metadata Schemas

Existing metadata schemas like Dublin Core lacks ETD-specific elements

Existing ETD schemas are insufficient to fulfill all aspects of FAIR compliance

3 major gaps

1

Incomplete representation

- “dc.rights” in ETD-MS v1.1 doesn’t support rich nuances of the rights information
    - only allows three values regarding accessibility
- “dc.format” assumes a single file format; ETDs often have multiple formats

2

Lack of elements describing parts of ETDs

Example: chapters, figures, tables

3

Lack of metadata elements added from sources other than those responsible for handling the ETD submission (e.g., authors, advisors, library catalogers)

- Needs provenance information for such metadata
- For example: For example, if we generate a chapter summary, the metadata must reference the process or model used

Need a More Descriptive Schema

1

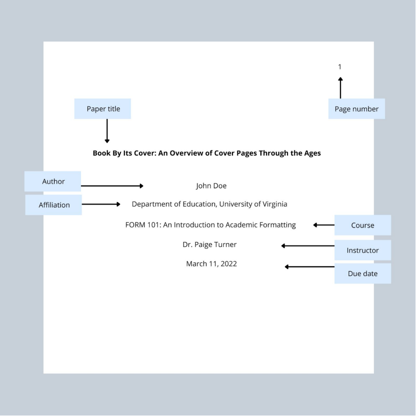
Capture detailed document-level metadata for ETDs

2

Fine-granular metadata standard to comply with FAIR principles

3

Capture detailed content-level metadata for ETDs



Book By Its Cover: An Overview of Cover Pages Through the Ages

John Doe

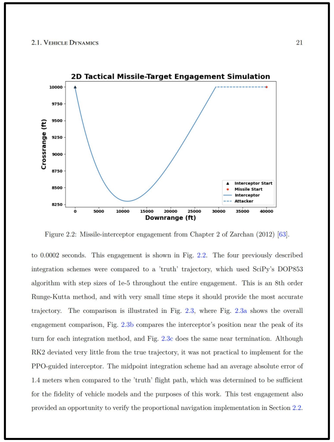
Department of Education, University of Virginia

FORM 101: An Introduction to Academic Formatting

Dr. Paige Turner

March 11, 2022

Document-level metadata - https://getproofed.com.au/writing-tips/how-to-format-an-apa-title-page/



2.1. Vehicle Dynamics 21

Figure 2.2: Missile-interceptor engagement from Chapter 2 of Zarchan (2012) [63].

to 0.0002 seconds. This engagement is shown in Fig. 2.2. The four previously described integration schemes were compared to a 'truth' trajectory, which used SciPy's DOP853 algorithm with step sizes of 1e-5 throughout the entire engagement. This is an 8th order Runge-Kutta method, and with very small time steps it should provide the most accurate trajectory. The comparison is illustrated in Fig. 2.3, where Fig. 2.3a shows the overall engagement comparison, Fig. 2.3b compares the interceptor's position near the peak of its turn for each integration method, and Fig. 2.3c does the same near termination. Although RK2 deviated very little from the true trajectory, it was not practical to implement for the PPO-guided interceptor. The midpoint integration scheme had an average absolute error of 1.4 meters when compared to the 'truth' flight path, which was determined to be sufficient for the fidelity of vehicle models and the purposes of this work. This test engagement also provided an opportunity to verify the proportional navigation implementation in Section 2.2.

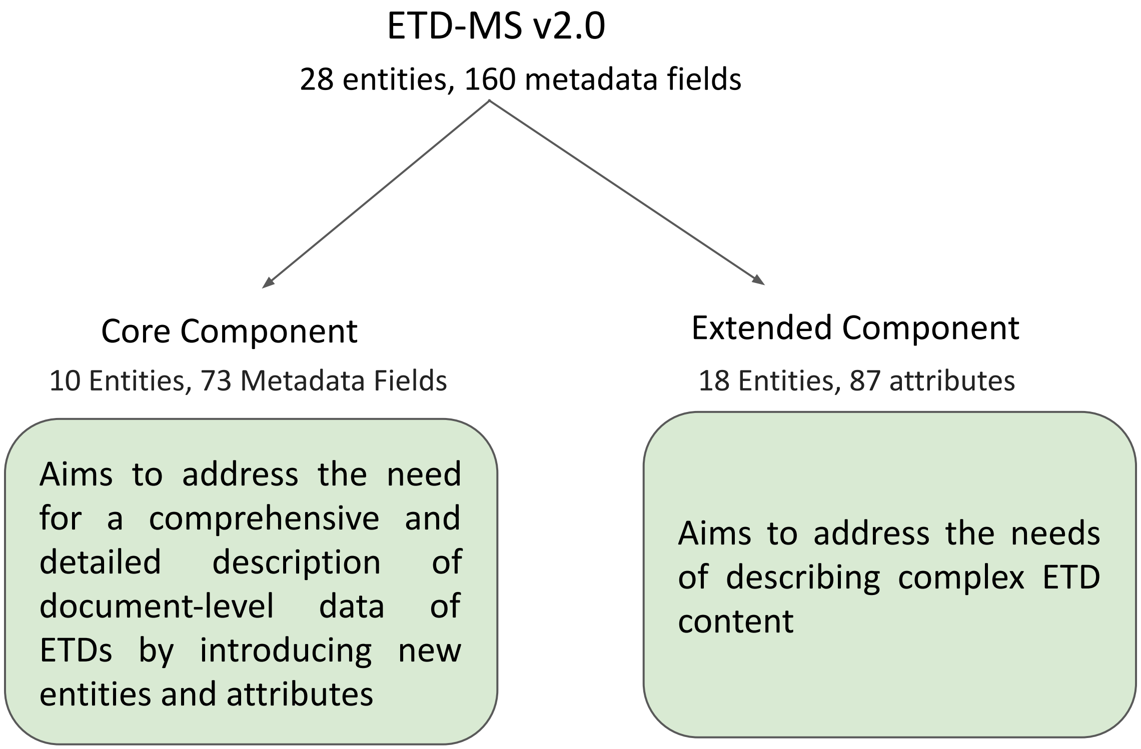
Metadata that describes the parts of an ETD, including various types of objects, their attributes and provenance

Content-level metadata - https://vtechworks.lib.vt.edu/server/api/core/bitstreams/b95e2ae9-4cf8-41e4-86cc-2148099547a3/content

ETD-MS v2.0: Bridging the Gaps in ETD Metadata

- Extends the existing metadata standard
- Core Component provides a comprehensive and detailed representation of document-level metadata
    - New entities to describe the ETD document itself and its relationships with other documents
        - For example:
            - Entities: “Rights”, “ETD_File”, “References”
    - Attribute for various abstracts
        - “abstractgeneral” for General Abstract
- Extended Component captures detailed content-level metadata
    - Metadata elements for describing parts of an ETD, such as chapters, figures, and tables
    - Provenance information for metadata
- Enhances *findability* and *reusability* of ETD at document and content levels

ETD-MS v2.0 Overview



ETD-MS v2.0

28 entities, 160 metadata fields

Core Component

10 Entities, 73 Metadata Fields

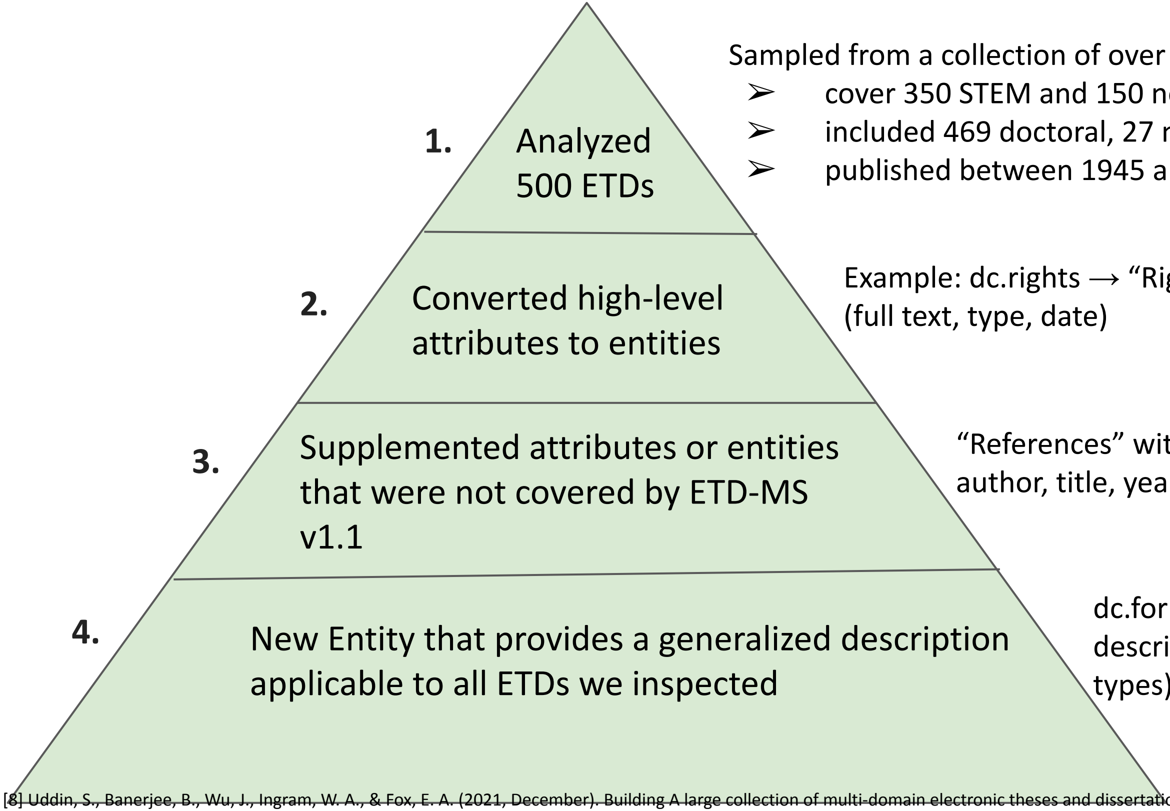
Aims to address the need for a comprehensive and detailed description of document-level data of ETDs by introducing new entities and attributes

Extended Component

18 Entities, 87 attributes

Aims to address the needs of describing complex ETD content

Core Component Development



1. Analyzed 500 ETDs

Sampled from a collection of over 500K ETDs from 114 U.S **[8]**

- ➢ cover 350 STEM and 150 non-STEM majors
- ➢ included 469 doctoral, 27 master’s, and 5 bachelor’s degrees
- ➢ published between 1945 and 1990

2. Converted high-level attributes to entities

Example: dc.rights → “Rights” entity with attributes (full text, type, date)

3. Supplemented attributes or entities that were not covered by ETD-MS v1.1

“References” with attributes such as reference_text, author, title, year, venue

4. New Entity that provides a generalized description applicable to all ETDs we inspected

dc.format (ETD-MS v1.1) → “ETD_File” entity (file description, generation method, checksum, MIME types)

[8] Uddin, S., Banerjee, B., Wu, J., Ingram, W. A., & Fox, E. A. (2021, December). Building A large collection of multi-domain electronic theses and dissertations. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 6043-6045). IEEE.

Core Component Description

10 Entities, 73 Metadata Fields, 2 Entity Categories

Category C.1

Describing the ETD document

Entities: “ETDs”, “Rights,” “ETD_file,” “Subjects,” “ETD_classes,” “ETD_topics”

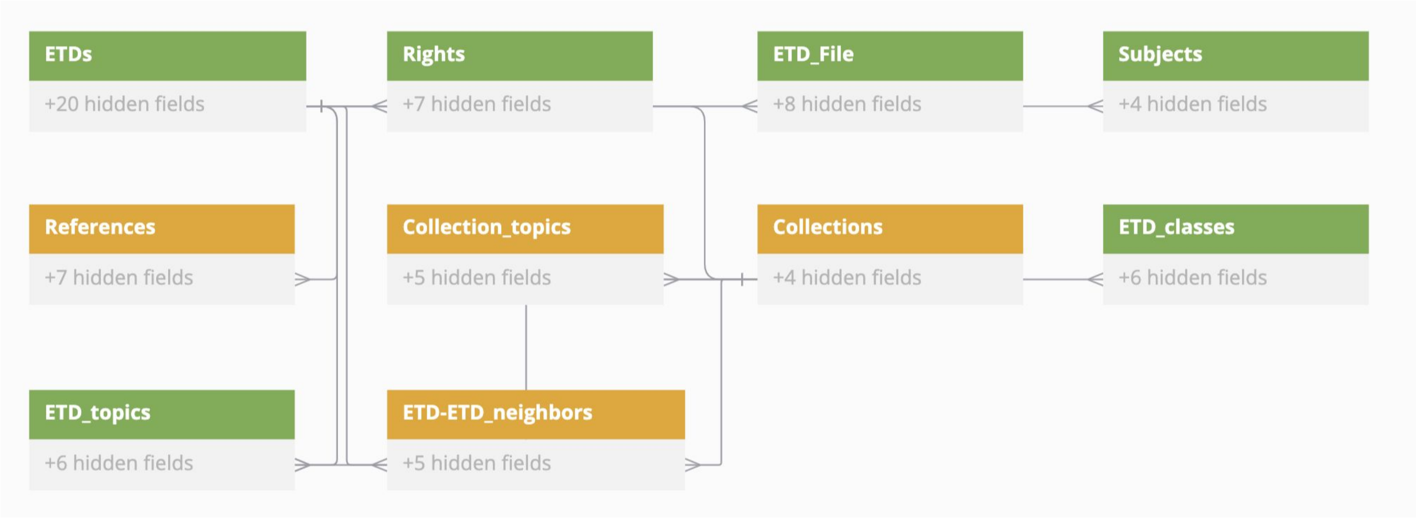
Category C.2

Describing the ETD’s relationship with other documents

Entities: “References,” “ETD-ETD_neighbors,” “Collections,” “Collection_topics”

Core Component Entities, Attributes, and Relationships

Figure 3: Relationships among Entities in the Core Components of ETD-MS v2.0.



🟩 Category C.1

🟧 Category C.2

Extended Component Development

- Bootstrap Approach
- Identified objects (e.g., chapters, figures, tables) and their attributes (e.g., text, object classes, object summaries).
- Added entities to describe object provenance and relationships
- Examples:
    - "Classifications": Classification systems (e.g., ProQuest Subject Categories)
    - "Classification_entries": Specific subject categories (e.g., "Library Science", "Web Studies")
    - "Classifiers": Models to classify ETDs or objects relative to a classification
    - User Interactions: Entities like "Users", "User_classes", and "User-user_neighbors"

Extended Component Description

Represent objects within ETDs, their provenance, relations, and user interactions
18 Entities, 87 attributes, 3 Entity Categories

Category C.1

Describing objects

Entities: “Object,” “Text,” “Object_classes,” “Object_summaries,” “Object_metadata,” “Object_topics”

Category C.2

Describing provenance

Entities: “Classifications,” “Classification_entries,” “Classifiers,” “Topic_models,” “Summarizers”

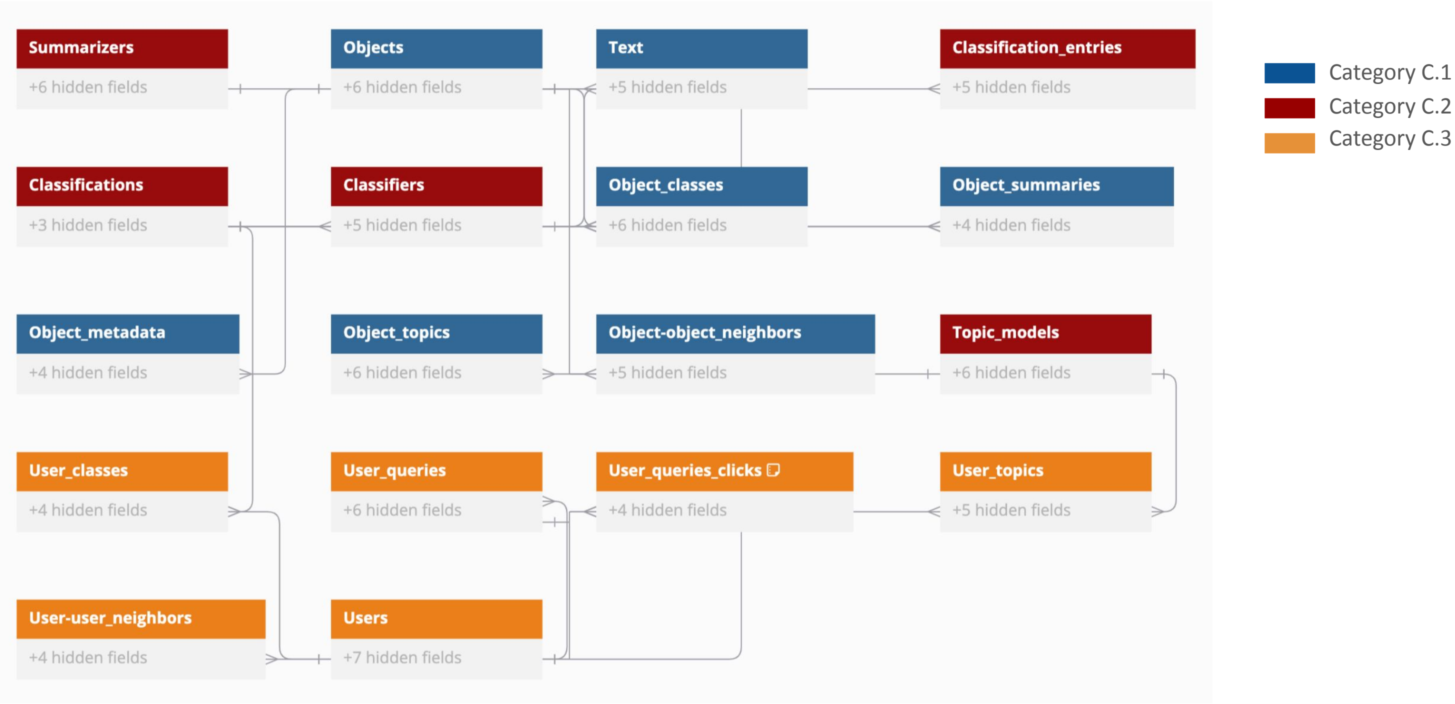
Category C.3

Describing user interactions

Entities: “Users,” “User_classes,” “User_queries,” “User_queries_clicks,” “User_topics,” “User-user_neighbors”

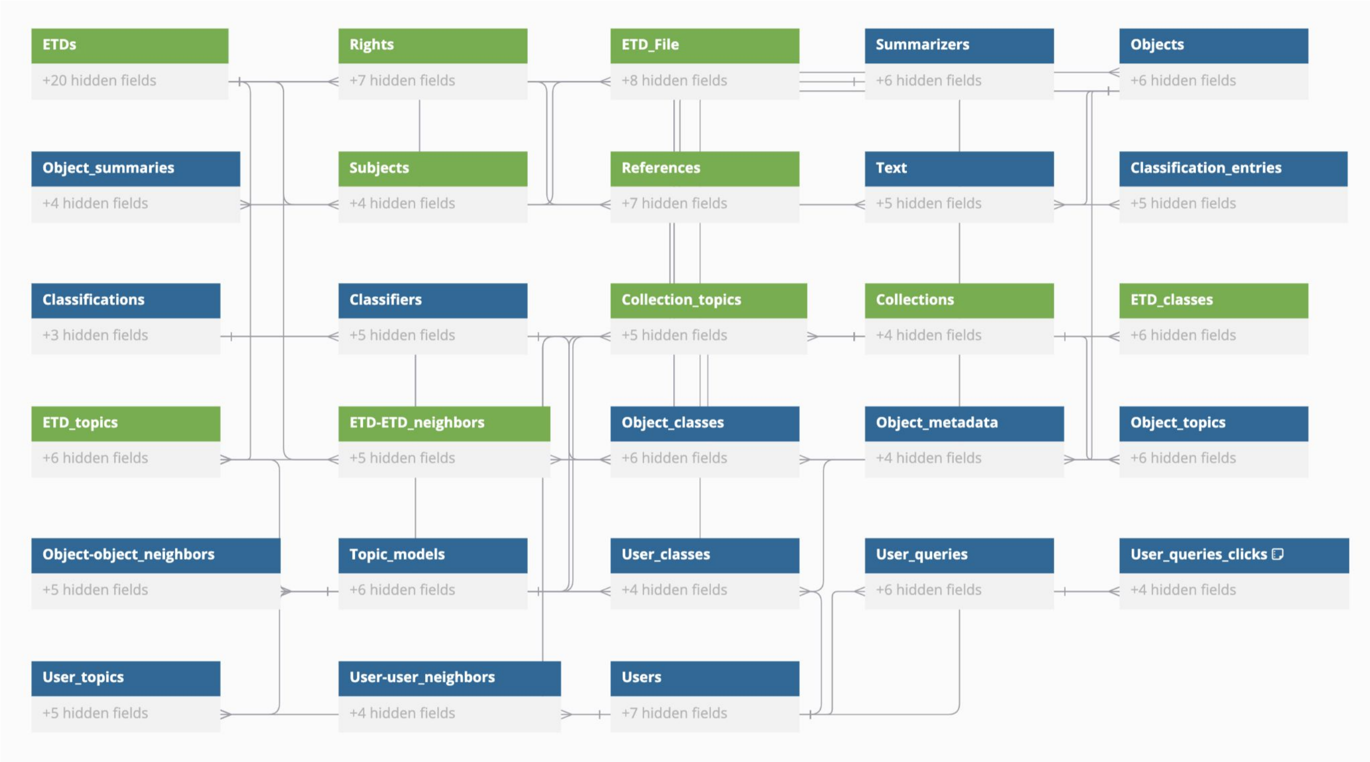
Extended Component Entities, Attributes, and Relationships

Figure 4: Relationships among Entities in the Extended Components of ETD-MS v2.0.



Entity Relationships in Core and Extended Components

Figure 5: Relationships among Entities in the Core and Extended Components of ETD-MS v2.0.



🟩 Core Component

🟦 Extended Component

Proof of Concept Database Implementation for Evaluation

Step 1: Database Design

- MySQL for database
- 28 tables each mapping to an entity in the ETD-MS v2.0 schema

Step 2: Data Population

- Selected 1,000 ETDs from a collection of over 500K ETDs
  - 50 U.S. universities, Year: 2005-2019
  - No overlaps with the 500 ETDs used for schema development
- Extracted 17 metadata fields (e.g., title, year, author)
- Scraped HTML files for additional metadata such as copyright details
- AI-based methods to derive content-level metadata
  - Classification: GPT-3.5 for ProQuest subject categories
  - Summarization: Fine-tuned T5-Small [9] and Pegasus [10] models
- NULL values for the unavailable metadata
- Dummy data for user-related tables

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... &amp; Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.
[10] Zhang, J., Zhao, Y., Saleh, M., &amp; Liu, P. (2020, November). Pegasus: Pre-training with extracted gap- sentences for abstractive summarization. In international conference on machine learning (pp. 11328-11339). PMLR.

Proof of Concept Database Implementation for Evaluation

- Performance and Scalability
  - Populated 1,000 ETD entries in ~11 minutes on a virtual machine (32 CPUs, 125 GB RAM)
  - Results demonstrated reasonable scalability
  - Most metadata fields successfully map to our new schema
    - Missing fields like “Peer-reviewed”
    - Missing different dates for ETD addition and accessibility

ETD-MS v2.0 Enhances FAIR Principles

Finability	Accessibility	Interoperability	Reusability
Improves ETD data findability through content-level metadata in digital library search engines	Improves accessibility through detailed rights information and ETD format	Interoperability is slightly reduced by new fields.	Reusability through collection subsets used for individual projects
Entities: “ETD_topics,” “ETD_classes,” “Object_metadata”	Entities: “Rights,” “ETD_File”	Mitigate this issue by mapping fields in the new schema to equivalent fields in the existing metadata schema. .	Entities: “Collections,” “Collection_topics,”

Example Mappings Between Metadata Schemas

Table 1: Mapping Between Simplified Dublin Core, ETD-MS v1.1, and the Core Component of ETD-MS v2.0.

Simplified Dublin Core	ETD-MS v1.1	ETD-MS v2.0
dc.title	dc.title	ETDs.title
dc.creator	dc.creator	ETDs.author
dc.description	dc.description.abstract	ETDs.abstract
	thesis.degree.grantor	ETDs.institution
		Rights.rights_uri

Limitations and Future Work

- Limitations
  - Based on 500 ETDs; limited scope
  - Missing fields like “Peer-reviewed”
  - Missing different dates for ETD addition and accessibility
- Future Work
  - Incorporate more fields for comprehensiveness
  - Collect feedback from ETD users

Resources

- For detailed documentation, schema design, and mapping information, visit our repository: https://github.com/lamps-lab/ETDMiner/tree/master/ETD-MS-v2.0