Metadata Enrichment of Low-Resource Scientific Data Using LLM Agent

Manika Lamba[1], You Peng[2], Sophie Nikolov[2], Glen Layne-Worthey[2], J. Stephen Downie[2]

[1]University of Oklahoma

[2]University of Illinois Urbana-Champaign

1. Introduction

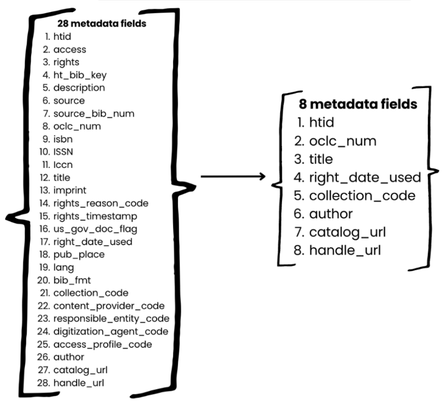
- Low-resource collections suffer from missing or inconsistent metadata
- Gaps reduce discoverability, accessibility, and long-term reuse
- Existing enrichment approaches are either manual or automated
- We propose a hybrid workflow for dissertation metadata enrichment in HathiTrust Digital Library (HTDL)

2. Data

- HTDL houses 19.3M+ digitized items (primarily books); dissertations remain underrepresented due to mass digitization gaps: incomplete metadata, uncorrected OCR, and collection bias [1, 2]
- Collected 6,445 English dissertation records from HTDL (1920–2011), with full text and metadata
- 1,403 records excluded due to mislabeling, bad OCR, or missing full-text
- Represented each record using 8 core metadata fields

Data Extraction & Cleaning

Category	Frequency
Not ETD	1131
Bad OCR	216
Full-Text file was not available	56



3. Metadata Enhancement Steps I–II

We enhanced the metadata manually, using the Dublin Core metadata standard to rename the corresponding HTDL field names, and conducted multiple rounds of manual data cleaning and normalization, ensuring consistency, accuracy, and usability.

4. Metadata Enrichment Step III

We enriched the metadata by adding 10 new dissertation-specific descriptors for *advisor, committee chair, committee member, department, university, degree name, degree level, location, keywords,* and *abstract* that had not been included in the original metadata.

Hybrid Metadata Enrichment Framework

Steps I–IV show the enrichment procedure applied to HTDL dissertation metadata and correspond to Sections 3–5

I. Refine the existing metadata by correcting errors, standardizing formats

ETDs with Multiple Author Names: ETDs with multiple authors were split into distinct rows for clarity. For example, for the ETD, *"The Nature of Gas-Metal Electrodes"*, co-authored by Sidney J. French and Louis Kahlenberg, a duplicate row was created, differing only by author name.

ETDs with Multiple Degrees: ETDs with multiple degrees were separated into unique rows for each degree. For instance, Patrick Francis Quinn's ETD on *"The Fatalism of Herman Melville"* was listed under separate rows for "Bachelor of Arts" and "Master of Arts."

Multiple Authors earning Different Degrees for same ETD: A more complex case involved ETDs with multiple authors earning different degree types. For ETDs co-authored by individuals earning different degrees, unique rows were created for each author-degree pair. For example, *"The Effect of Decreased Oxygen in the Respired Air Upon Metabolism, Acidosis, and the Differential Leukocyte Count"* involved five authors with varying degrees (e.g., Master of Science, Master of Arts, Bachelor of Science in Medical Sciences).

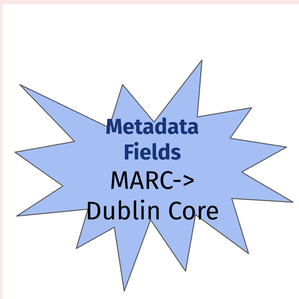
Author's Name Formatting: Periods were removed from first names (e.g., "Margaret Anne." became "Margaret Anne").

Date of Birth Standardization Birth years formatted with a leading hyphen (e.g., "-1887") were corrected to avoid confusion, resulting in entries like "1887."

Removing Duplicate and non-ETD Titles: We excluded 1,131 entries misclassified as ETDs, along with duplicate titles.

De-duplication Complexity: De-duplication was more complex than anticipated. There were several ETDs that were integrated back into the main metadata file because of version errors and discrepancies between the *'dc:identifier: catalog'* field versus the *'dc:identifier:handle'* field. In these cases, *'handle-url'* showed a version of the ETD without a title page whereas the *'catalog-url'* field for the same ETD, provided multiple versions of the ETD where at least one of those versions provided a title page that was conducive to adding more metadata.

II. Align data with established metadata schemas or standards



*Renamed the field names using **Dublin Core** metadata standard created by National Digital Library of Theses and Dissertations (NDLTD)*

Table 1: Metadata Enrichment for HathiTrust ETDs using Dublin Core Fields

Older Field Name	New Field Name	Description
htid	dc:identifier:htid	Unique ID number categorizing an ETD, characterized by a string of letters to symbolize the institution collection, followed by a series of numbers assigned for HathiTrust data
oclc-num	dc:identifier:oclc	OCLC (Online Computer Library Center) control number for the item
collection-code	dc:relation:isPartOf	Code identifying the specific collection within HathiTrust
title	dc:title	The title of the ETD
author	dc:creator	Name of the author who wrote the ETD
author	dc:creator:dob	Author's date of birth, if available, and date of death, if applicable
rights_date_used	dc:date:issued	Publication year
–	dc:contributor:advisor	Name of the faculty member who served as the primary advisor for the ETD
–	dc:contributor:committee Chair	Name of the faculty member who served as the chair of the ETD
–	dc:contributor:committee Member	Names of faculty members who served on the thesis/dissertation committee
–	thesis:degree:department	Academic department responsible for the thesis/dissertation program
–	thesis:degree:discipline	Academic field or discipline of study
–	thesis:degree:grantor	Institution awarding the degree
–	thesis:degree:name	Name of the degree awarded (e.g., Ph.D., M.S., Ed.D.)
–	thesis:degree:level	Level of the degree (e.g., Doctoral, Master's)
–	dc:coverage:spatial	Location of the university
–	dc:subject	Keywords or subject terms describing the thesis content
–	dc:description:abstract	Summary or abstract of the thesis/dissertation content
catalog-url	dc:identifier:catalog	Permanent URL or identifier for the catalog record
handle-url	dc:identifier:handle	The direct url link to the ETD provided by HathiTrust

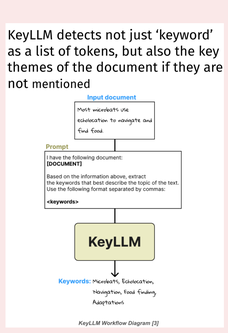
III. Extend beyond basic refinement and add new layers of information to metadata records to enhance context and relevance

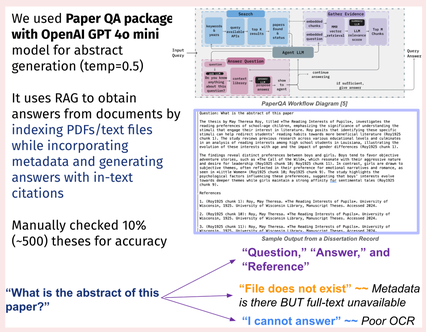
Added 10 more fields that we deemed crucial for describing the ETD data

Table 1: Metadata Enrichment for HathiTrust ETDs using Dublin Core Fields

Older Field Name	New Field Name	Description
htid	dc:identifier:htid	Unique ID number categorizing an ETD, characterized by a string of letters to symbolize the institution collection, followed by a series of numbers assigned for HathiTrust data
oclc-num	dc:identifier:oclc	OCLC (Online Computer Library Center) control number for the item
collection-code	dc:relation:isPartOf	Code identifying the specific collection within HathiTrust
title	dc:title	The title of the ETD
author	dc:creator	Name of the author who wrote the ETD
author	dc:creator:dob	Author's date of birth, if available, and date of death, if applicable
rights_date_used	dc:date:issued	Publication year
–	dc:contributor:advisor	Name of the faculty member who served as the primary advisor for the ETD
–	dc:contributor:committee Chair	Name of the faculty member who served as the chair of the ETD
–	dc:contributor:committee Member	Names of faculty members who served on the thesis/dissertation committee
–	thesis:degree:department	Academic department responsible for the thesis/dissertation program
–	thesis:degree:discipline	Academic field or discipline of study
–	thesis:degree:grantor	Institution awarding the degree
–	thesis:degree:name	Name of the degree awarded (e.g., Ph.D., M.S., Ed.D.)
–	thesis:degree:level	Level of the degree (e.g., Doctoral, Master's)
–	dc:coverage:spatial	Location of the university
–	dc:subject	Keywords or subject terms describing the thesis content
–	dc:description:abstract	Summary or abstract of the thesis/dissertation content
catalog-url	dc:identifier:catalog	Permanent URL or identifier for the catalog record
handle-url	dc:identifier:handle	The direct url link to the ETD provided by HathiTrust

IV. Add meaning and contextual connections to the document's content





(a) Keyword Extraction **(b)** Abstract Generation

5. Semantic Enrichment Step IV

Following the manual process, we then used an automated method to enrich the metadata semantically by using KeyLLM [3] for keyword extraction and PaperQA [5], a RAG-based LLM agent for text summarization, to populate the 'keywords' and 'abstracts' metadata fields that were not part of the original HTDL metadata representation.

6. Conclusion

This study presents a hybrid workflow for enriching low-resource dissertation metadata in HathiTrust Digital Library (HTDL). By combining manual metadata enhancement with LLM-based semantic enrichment, the approach introduces additional access points which may not have been originally designed to accommodate other types of content. This work can be adapted to other digital library collections with incomplete or inconsistent metadata.

Dataset Availability

The dataset introduced and analyzed in this paper is available for download and reuse under a CC BY-NC-SA 4.0 license in Zenodo. A citation to the dataset is provided in the reference list, as Lamba et al. [4].

References

[1] Katherine Bode. Why You Can't Model Away Bias. *Modern Language Quarterly*, 81(1):95–124, 2020.

[2] Nicole M. Brown et al. In Search of Zora/When Metadata Isn't Enough: Rescuing the Experiences of Black Women Through Statistical Modeling. *Journal of Library Metadata*, 19(3-4):141–162, 2019. https://doi.org/10.1080/19386389.2019.1652967.

[3] Maarten Grootendorst. Introducing KeyLLM — Keyword Extraction with LLMs, 2024.

[4] Manika Lamba et al. Metadata Enrichment of Long Text Documents using Large Language Models. *Proceedings of the Association for Information Science and Technology*, 62(1):990–994, 2025.

[5] Jakub Lála et al. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research, 2023. arXiv:2312.07559.