MetaEnhance: Metadata Quality Improvement for Electronic Theses and Dissertations

Muntabir Hasan Choudhury, Lamia Salsabil, Himarsha R. Jayanetti, Jian Wu
Department of Computer Science
Old Dominion University

Introduction

- Metadata quality is important for digital objects to be discovered through digital library (DL) systems.
- This metadata often exhibits incomplete, inconsistent, and incorrect values [1].
- Existing frameworks rely on a semi-automatic approach or manual corrections such as crowdsourcing, e.g., [2]. Such methods are slow and biased toward document discoverability.
- Took electronic theses and dissertations (ETDs) as a case study where we found poor metadata quality. It may harm the discoverability of DL.
- Proposed MetaEnhance, a framework that utilizes state-of-the-art artificial intelligence (AI) methods to improve the quality of metadata on a benchmark dataset containing 500 ETDs.
- The framework achieved a remarkable performance by automatically detecting, correcting, and canonicalizing the surface values of seven key metadata fields, including title, author, university, year, degree, advisor, and department, which are ubiquitous in ETDs.

Methodology

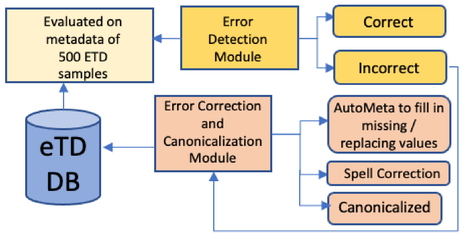

Figure 1: MetaEnhance Framework

Field	Acronym	Full Name
University	OSU	Ohio State University
Degree	MS, M.S.	Master of Science
Department	MSE, MSCE	Materials Science and Engineering

Table 1: Acronym and Full name example

MetaEnhance (Figure 1) is comprised of three main modules: error detection, correction, and canonicalization. First, it detects three types of errors – missing values, incorrect values, and misspellings. Then, the correction module corrects, canonicalize, and utilizes AutoMeta [3] to fill in missing values and overwrite any incorrect metadata by directly extracting the metadata from ETDs.

Error Detection (ED) and Error Correction & Canonicalization (ECC)

- **Title** found error titles, such as DMA Recitals. Adopted the title classifier by Rohatgi et al. [4] to detect errors. Any incorrect title was overwritten with AutoMeta [3] result.
- **Author and Advisor** adopted FlairNLP (named entity recognition model). If the author/advisor name is classified as a type other than PERSON, considered as incorrect and was overwritten with AutoMeta [3] result.
- **University** found only acronyms (*Table 1*). Built a dictionary-based method to classify acronyms as an error. If acronym exists, the correction module replaces acronyms with their full names.
- **Degree** found acronyms (*Table 1*) and incorrect metadata, e.g., history. Built dictionary-based method consisting of degree full name and acronyms. If acronym exists, the correction module replaces acronyms with their full names.
- **Department** found numerous misspellings (e.g., scool of Music). Used pyspellchecker to detect and correct errors. Also, found different forms of the same department names (*Table 1*). To disambiguate the surface name, the module used SentenceTransformer and a dictionary-based method.
- **Year** found inconsistent format across libraries, such as: mm-dd-yyyy or yyyy-mm-dd. Used dateutil.parser.parse and dictionary-based method to detect errors. If an error exists, overwrite the value from AutoMeta [3]. To canonicalize, used regular expressions and dateutil.parser.parse to parse and store year, month, and date in three separate columns.

Evaluation & Results

- Our benchmark dataset contains metadata from 500 ETDs *(Table 2)* selected from 533,047 ETDs by crawling 114 US university libraries and ProQuest.
- Evaluated on this benchmark and reported precision, recall, and F1 scores.
- *Table 3* shows that ED module achieved F1 > 0.99 for the title, author, and department fields and perfect recall and precision for university, year, and degree fields.

Field	#Missing	#Canonical	#Spell	#Incorrect
Title	0	0	0	1
Author	2	0	0	0
Advisor	150	35	0	0
University	6	43	0	0
Year	172	1	0	0
Degree	156	82	0	4
Department	269	85	2	0

Table 2: Distribution of ETD errors in the 500 benchmark dataset

- The performance of ECC relies on the output of AutoMeta [3].
- Except title and author, ECC successfully corrected all missing values. Also, the module successfully canonicalized the surface names.
- ECC canonicalized 7%, 6.4%, 0.2%, 16.2%, and 16.6% for the advisor, university, year, degree, and department fields, respectively.
- *Table 3* shows that ECC successfully corrected 4 incorrect values, and 2 misspellings, for the degree and department fields, respectively.

Field	P_{ED}	R_{ED}	$F1_{ED}$	P_{ECC}	R_{ECC}	$F1_{ECC}$
Title	0.997	1.0	0.998	0.0	0.0	0.0
Author	0.996	1.0	0.997	0.0	0.0	0.0
Degree	1.0	1.0	1.0	0.980	1.0	0.980
Department	0.996	1.0	0.997	0.970	1.0	0.980
University	1.0	1.0	1.0	0.740	1.0	0.850
Year	1.0	1.0	1.0	1.0	1.0	1.0
Advisor	0.920	0.990	0.950	1.0	1.0	1.0

Table 3: Error Detection & Correction Evaluation

Error Analysis – Found one false positive (FP) and one false negative (FN) for the title. For example, DMA Recitals was misclassified as valid. Also found 2 FPs for the author field. For the advisor field, the ED achieved F1=0.95 with 37 FPs. The department classifier correctly detected 2 misspellings (e.g., scool in the department field) but misclassified 2 surface values. For example, Public Health (PMH) was classified as an incorrect department name.

Conclusion and Discussion

- MetaEnhance demonstrated effectiveness in automatically improving ETD metadata quality when evaluated on 500 benchmark datasets.
- Achieved 95%–99% F1 in detecting errors and corrected 85%–98% (F1).
- **Limitation** – a dictionary-based approach for university, degree, and department to detect errors may misclassify if not found in dictionary.

References

[1] Bui and Park. An assessment of metadata quality: A case study of the national science digital library metadata repository. 2013.

[2] Wu et al. The impact of user corrections on a crawl-based digital library: A citeseerx perspective. 2014.

[3] Choudhury et al. Automatic metadata extraction incorporating visual features from scanned electronic theses and dissertations. 2021.

[4] Rohatgi et al. What were people searching for? a query log analysis of an academic search engine. 2021.