# MetaEnhance: Metadata Quality Improvement for Electronic Theses and Dissertations of University Libraries

Muntabir Hasan Choudhury
Lamia Salsabil
Old Dominion University
Norfolk, United States
{mchou001,lsals002}@odu.edu

Himarsha R. Jayanetti
Jian Wu
Old Dominion University
Norfolk, United States
{hjaya002,j1wu}@odu.edu

William A. Ingram
Edward A. Fox
Virginia Tech
Blacksburg, United States
{waingram,fox}@vt.edu

## Abstract

Metadata quality is crucial for discovering digital objects through digital library (DL) interfaces. However, due to various reasons, the metadata of digital objects often exhibits incomplete, inconsistent, and incorrect values. We investigate methods to automatically detect, correct, and canonicalize scholarly metadata, using seven key fields of electronic theses and dissertations (ETDs) as a case study. We propose MetaEnhance, a framework that utilizes state-of-the-art artificial intelligence (AI) methods to improve the quality of these fields. To evaluate MetaEnhance, we compiled a metadata quality evaluation benchmark containing 500 ETDs, by combining subsets sampled using multiple criteria. We evaluated MetaEnhance against this benchmark and found that the proposed methods achieved nearly perfect F1-scores in detecting errors and F1-scores ranging from 0.85 to 1.00 for correcting five of seven key metadata fields. The codes and data are publicly available on GitHub[1].

## CCS Concepts

• **Applied computing** → **Digital libraries and archives**; **Document metadata**; • **Information systems** → **Incomplete data**; **Inconsistent data**; • **Computing methodologies** → **Machine learning**.

## Keywords

Digital Libraries, Scholarly Big Data, ETD, Metadata Quality, Artificial Intelligence

## 1 Introduction

Metadata represents a key aspect of digital objects. Improving metadata quality for DL objects is a long-standing problem. Although DL systems have adopted Dublin Core (DC) to standardize metadata formats (e.g., ETD-MS v1.1), studies have shown frequent inaccurate, incomplete, and inconsistent metadata elements in DLs [2]. To address metadata quality issues, in one survey paper [9], the authors discussed the overlaps of quality assessment frameworks and defined metadata quality parameters, dimensions, and metrics. One existing method to improve DL metadata is crowdsourcing, letting users manually correct metadata errors [11]. This method has two drawbacks – a) it is difficult to control the user population, and b) it is slow and thus not scalable. Most existing frameworks rely on semi-automatic approaches or manual corrections. With the advancement of AI, it is possible to explore natural language processing (NLP) and computer vision (CV) methods to improve metadata quality by automatically detecting, correcting, and canonicalizing metadata. Due to the heterogeneous nature of digital objects, designing a single system that fixes all metadata fields is challenging. Here, we use ETD metadata as a case study.

ETDs are scholarly documents that represent students' research and demonstrate their ability to independently conduct and communicate research findings and meet the requirements for an academic degree. ETDs are usually hosted by university libraries or centralized online repositories such as ProQuest. The metadata of ETDs was originally input into the system by students, faculty, or library staff. Presumably, they should be complete, consistent, and accurate. However, upon inspecting metadata downloaded from several university libraries, we found many ETD repositories are accompanied by incomplete, inconsistent, and incorrect metadata. As reported in a paper [10], at least 43% of department and 12% of year fields were empty. Low metadata quality may significantly harm the discoverability of digital libraries.

In this paper, we propose a framework called MetaEnhance, aiming at improving ETD metadata quality by automatically filling in missing values, detecting and correcting errors, and canonicalizing surface names. We quantitatively demonstrated the effectiveness of our system in improving the seven key metadata fields, including title, author, university, year, degree, advisor, and department in ETDs. Our contributions are the following. a) We proposed MetaEnhance to improve the metadata of ETDs using AI methods. b) We created a new benchmark dataset using real-world ETD metadata to evaluate metadata quality improvement methods. c) Our proposed framework achieved a remarkable performance to improve metadata quality in the benchmark data.

## 2 Related Work

Several digital libraries allow users to manually correct metadata. For example, when Microsoft Academic was online, it allowed users to change header information, including titles, authors, year, DOI, conference, journal, URL, and abstract. Wu et al. proposed user corrections as a form of crowd collaboration, providing an efficient way to improve metadata quality for CiteSeerX [11]. The authors inspected the correction history and showed that user correction was a reliable source of high-quality metadata. However, the paper only examined authors and titles.

Park et al. [6] argued the existence of inconsistent metadata because many different data providers may not strictly follow the DC schema. The author discussed and compared several methods to measure metadata quality and emphasized the most commonly used criteria, including accuracy, completeness, and consistency. The author compared published methods that proposed guidelines, best practices, and approaches for quality assurance. The author

[1] https://github.com/lamps-lab/ETDMiner/tree/master/metadata_correction

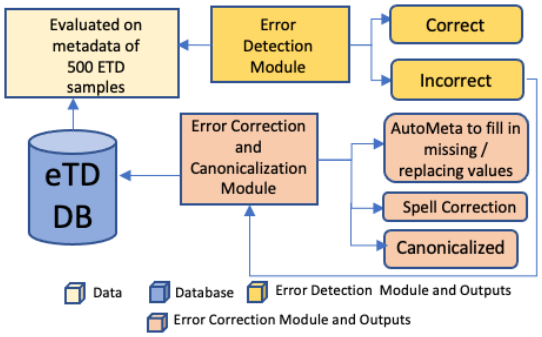


**Figure 1: MetaEnhance Framework.**

advocated the development of a framework for assessing quality and mechanisms to improve metadata quality. However, to our best knowledge, no AI-based frameworks have been proposed and implemented to improve metadata quality for ETDs.

In our proposed work, we focus on using AI methods and models to *automatically* improve metadata quality, which is more scalable than manual approaches.

## 3 Methodology

Figure 1 illustrates the framework, which is comprised of three main modules: error detection, correction, and canonicalization. We describe each module in the following sections.

### 3.1 Error Detection

Metadata usually has different data types and meanings. We detect three types of errors: missing values, incorrect values, and misspellings. If missing values are detected, the module switches directly to the correction module, in which the metadata is parsed from the ETD document using a machine-learning method that was previously developed and validated. We then inserted the metadata into the corresponding field. If the field is not empty, the module checks whether the content contains errors. Below, we elaborate on the error detection methods. Each field in our study corresponds to the field in the DC (in parentheses after field names)[2].

***Title*** *(dc.title)* This field might contain incorrect ETD titles, such as, "DMA Recitals", whereas the correct title is, "Expanding Vision: The Music of Alyssa Morris". We adopted the classifier proposed by Rohatgi et al. [8] to automatically detect incorrect titles. Each string in the title field is represented by the following features: the number of tokens, the number of special characters, the number of capital letters, consecutive punctuation, stop words, and the minimum, maximum, and median TF-IDF. The classifier was evaluated on the SciDocs [5] dataset and achieved an F1 score of 0.96.

***Author and Advisor*** *(dc.creator and dc.contributor)* We adopted a named entity recognition model implemented in the FlairNLP [1] package to automatically detect incorrect author and advisor names. This model was pre-trained and evaluated on the *Ontonotes*[3] dataset and achieved an F1 of 0.90. An error is detected if the author or advisor name or any part of it is classified as a type other than *PERSON*.

[2]https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html
[3]https://catalog.ldc.upenn.edu/LDC2013T19

**Table 1: Examples of acronym/colloquial names and corresponding full names for university, degree, and department fields.**

| Field | Acronym/Colloquial | Full Name |
|---|---|---|
| University | JHU, jhu | Johns Hopkins University |
| Degree | MPHIL, M PHIL, PHM | Master of Philosophy |
| Department | MSE, MSCE | Materials Science and Engineering |

***University*** *(thesis.degree.generator)* The university field contains different surface names referring to the same university. For example, "Johns Hopkins University" is abbreviated as "jhu" or "JHU." We built a dictionary[4] that consists of 832 names of universities of the United States and their acronyms. We checked the surface values against this dictionary and determined if any value was incorrect. If the field was marked as an error, the correction module will inspect it and decide whether it should be corrected or canonicalized.

***Degree*** *(thesis.degree.name)* For the degree field, our database record shows both acronyms and incorrect metadata. For example, the word "history" may appear in this field. We used a dictionary-based method [4] by compiling a dictionary containing 234 degree names and their acronyms based on the degree naming convention of *DegreeAbbreviations*[5][6]. Any value for a degree not found in this dictionary was identified as an error. However, these degree names may not be "wrong". They are marked as "errors" so further modules will inspect whether they are errors or need to be canonicalized.

***Department*** *(thesis.degree.discipline)* We observed numerous misspellings for department fields such as "College of Muisc", "scool of Music", "Graduhte Studies in English". To detect spelling errors in surface names, we used the Python library *pyspellchecker*[7] which uses Levenshtein distance to detect spelling errors. If the editing distance between the original word and the field value is 2, it is considered a spelling error.

***Year*** *(dc.date.issued)* The format of this field is inconsistent across libraries, such as: "mm-dd-yyyy" or "yyyy-mm-dd". We used the Pandas built-in parser "to_datetime" [8], which can perform generic parsing of dates in almost any string. We verified if the specific date field was valid using this parser and then checked against a dictionary listing the year range from 1880 to 2023. If we marked the value as an error, the correction module could inspect it and decide whether it should be corrected or canonicalized.

### 3.2 Error Correction & Canonicalization (ECC)

The ECC module contains two types of corrections depending on the errors detected in the previous model and in addition to canonicalizing entity surface names. These corrections are: a) filling in missing values using AutoMeta [3] and b) correcting misspellings and incorrect values.

[4]https://en.wikipedia.org/wiki/List_of_colloquial_names_for_universities_and_colleges_in_the_United_States
[5]https://abbreviations.yourdictionary.com/articles/degree-abbreviations.html
[6]https://degree.studentnews.eu/
[7]https://pypi.org/project/pyspellchecker/
[8]https://pandas.pydata.org/docs/reference/api/pandas.to_datetime.html

One challenge of the metadata correction task is the information source that can be used for filling in missing values and overwriting error values. The MetaEnhance system integrates an existing framework called **AutoMeta** [3], a framework that automatically extracts seven key metadata fields by combining visual and text features from ETD cover pages using conditional random fields (CRFs). The model was tested on a corpus of 500 ETDs (different from this paper) and achieved 81.3%–96% F1 scores depending on the field. We used the best model to extract metadata from ETDs used in this paper and fill in missing values for seven metadata fields. **Canonicalization** involves converting data with multiple possible surface names into a "standard" form. We canonicalized the values of the advisor, university, department, degree, and year fields detected by the previous model and the AutoMeta [3] result.

***Title*** For any title detected as an error, we overwrite it with the title extracted by the AutoMeta [3] for that specific ETD.

***Author and Advisor*** If an author or an advisor name was detected as an error, we overwrite the field using corresponding fields extracted using AutoMeta [3]. We also observed that advisor names, such as, "Mark Pankow, Co-Chair" or "Andrew Mathew Jr., Committee Member" needs to be parsed. According to DC, Co-Chair is a role *(dc.contributor.role)* of a member in the thesis committee. We used regular expressions to parse the surface value and then stored the value in a separate column.

***University*** If the university name was detected as an error, we employed a dictionary-based method by matching a university name against the university dictionary. We normalized both the field name and the colloquial names by converting all letters to upper case, stripping off punctuation marks, and then searching for the surface name in the dictionary to see if it was colloquial. The full name was then used to replace the surface name in this field. Table 1 shows examples of colloquial and official names. If any incorrect university is detected, we overwrite the incorrect university with the title extracted by AutoMeta [3].

***Degree*** If an error was detected in the degree field, we attempt to canonicalize the values by searching it against the *DegreeAbbreviations* dictionary. We first normalized field values that involved converting degree metadata to uppercase letters and removing all punctuation marks. We then replaced acronyms with their full forms. If any incorrect degree is detected, we overwrite the incorrect degree with the degree extracted by AutoMeta [3].

***Department*** The Department names can have different forms. For example, "Dept of CS" and "CS Department" all map to the same entity. For this field, we correct spelling errors, disambiguate department names, and canonicalize them into full names. The Python library *pyspellchecker*[6] was used to identify and correct **spelling errors**. The library captures errors using Levenshtein distance (see Section 3.1). All permutations are compared to known words in a word frequency list. Words that appear more often are considered correct spelling. To **canonicalize** department names, we compiled a comprehensive list of 232 different academic department names and their acronyms using the official *Abbreviations and Symbols from Boston University*[9]. We normalized all the surface names and the acronyms on the list. These surface names were compared with the acronyms on the list to see if they were abbreviated forms. If a match was found, the matching surface name was replaced with its corresponding full name from the list. To **disambiguate** department names, we developed a model using SentenceTransformers [7], which first generates embeddings of surface names and the department full names and then measures cosine similarity of the surface names against the dictionary set. We observed that 91% of the records with $cosine-similarity \geq 0.90$ provide correct matches. Table 1 shows an example that our model identifies the similarity and maps them to full names.

[9]https://www.bu.edu/academics/bulletin/abbreviations-and-symbols/

**Table 2: Distribution of ETD errors in the dataset used in this paper. The #Canonical column shows the count of values that needs canonicalization for each field. This count includes the values after missing fields are inserted and errors are corrected. The title and author fields do not contain surface names that should be canonicalized.**

| 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 |

***Year*** If an error was found in the year field, we used the value from AutoMeta [3]. To canonicalize the surface values, we utilized the Pandas built-in parser "to_datetime" (see Section 3.1), which outputs three date fields, including "year", "month", and "date" and stored them in three separate columns.

## 4 Evaluation and Results

To evaluate MetaEnhance, we compiled a corpus containing metadata from 500 ETDs selected from 533,047 ETDs crawled from 114 US university libraries, including full text in PDF format and metadata from university library repositories and ProQuest. Data collection was based on a software framework, which harvested ETDs and their metadata via the Open Archives Initiative protocol (OAI-PMH) or sitemaps [10]. To mitigate selection bias against samples that are under-represented in certain dimensions (e.g., random sampling would be biased against ETDs of minority universities), we selected 4 ETD subsets based on 4 different criteria and then combined them. For each criterion, except for the title and author fields, we selected ETDs with missing values for the remaining fields. The selection criteria ensure that we cover samples with errors in different metadata fields. Table 2 illustrates the distributions of ETDs in different feature spaces.

- **Random:** We randomly sampled 100 ETDs from our collection.
- **University:** We first randomly selected 10 universities and then 10 random ETDs from each university.
- **Year:** We randomly sampled 10 ETDs each year from 2010 – 2019.
- **Department:** We randomly sampled 6 STEM and 4 non-STEM disciplines. Then we randomly sampled 10 ETDs of each category.
- **Degree:** We randomly sampled 5 degree names and then selected 20 ETDs from each degree.

### 4.1 Error Detection Evaluation

Errors include missing values, misspellings, and incorrect values (e.g., titles). Table 2 shows the number of missing values for each field. Our error detection module correctly detected all the missing values for each field. Depending on the metadata types, the module detects errors differently. We checked against the ground truth for each field and reported precision, recall, and F1 scores. Table 3 shows that the university, year, and degree fields achieved perfect recall and precision. The error detection module achieves $F1 > 0.99$ for the title, author, and department fields.

We found one false positive (FP) and one false negative (FN) for the title field. For example, "DMA Recitals" was misclassified as a valid title. We also found 2 FPs for the author field. The classifier misclassified the name "Richmond Orien Manu Wright". Here, the first name "Richmond" was identified as a geopolitical entity (*GPE*), while the rest was identified as a *PERSON*. For the advisor field, our method achieved 0.95 F1 with 37 FPs. Specifically, the method misclassified the name "Mark Pankow, Co-Chair". While the department classifier correctly detected 2 misspellings (e.g., "scool"), it misclassified 2 surface values. For example, "Public Health (PMH)" was classified as an incorrect department name.

### 4.2 ECC Evaluation

The performance of the ECC module relies on the output of AutoMeta [3]. Although AutoMeta achieved F1-scores of 0.67 – 0.91 for most fields on the 500 evaluation benchmark, it failed to extract advisor field values from most of the ETDs, because most advisor fields do not appear on the cover pages of the ETDs in our corpus. Further, the extraction quality of AutoMeta depends on the scan resolution of ETDs.

The title field does not contain missing or incorrect values (Table 2). The error detection module only detected two missing values for the author field. However, AutoMeta could not extract authors for these two ETDs to fill in those missing values, which leads to zero precision, recall, and F1. For the other fields, the ECC module successfully corrected all missing values. Moreover, the module successfully canonicalized the surface names in degree (e.g., "Ph.D."), department (e.g., "CS"), and university ("JHU") fields. Table 2 shows the total number of values needing to be canonicalized. We canonicalized 7%, 8.6%, 0.2%, 16.2%, and 16.6% for the advisor, university, year, degree, and department fields, respectively. Table 3 shows that we successfully canonicalized all values of year and advisor fields and a high percentage of values in other fields. When the department and degree fields were incorrectly extracted by AutoMeta [3], they were not canonicalized. Furthermore, the ECC module successfully corrected 4 incorrect values, and 2 misspellings, for the degree and department fields, respectively.

## 5 Conclusion and Discussion

We developed MetaEnhance, a system to automatically improve ETD metadata using AI methods. We then applied it to our benchmark dataset and quantitatively demonstrated the effectiveness of our system in improving the metadata quality. Overall F1 scores, depending on the metadata fields, MetaEnhance achieved 95%–99% in detecting errors and corrected 85%–98% of errors, e.g., filling missing values and canonicalizing surface names. One limitation of the error detection module is that it may misclassify a valid university or degree if it is not found in the dictionary. In addition, the department field canonicalization only maps the acronyms to the full names when they are included in the Boston University dictionary list of department names. It is possible that universities may not follow exactly the same convention for certain acronyms. The benchmark data can be made more challenging by incorporating more spelling errors and incorrect values, which can be achieved by introducing random noise to true field values.

**Table 3: Performance of Error Detection (ED) and Error Correction and Canonicalization (ECC).**

| 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 |

## Acknowledgement

Support was provided by the Institute of Museum and Library Services through grant LG-37-19-0078-198. We thank Dominik Soos for collecting and ingesting ETDs from minority serving institutions.

## References

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1638–1649. https://aclanthology.org/C18-1139/

[2] Yen Bui and Jung-ran Park. 2006. An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository. *Proceedings of CAIS/ACSI* 2006 (01 2006). https://doi.org/10.29173/cais166

[3] M. Hasan Choudhury, H. R. Jayanetti, J. Wu, W. A. Ingram, and E. A. Fox. 2021. Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, Los Alamitos, CA, USA, 230–233. https://doi.org/10.1109/JCDL52503.2021.00066

[4] Muntabir Hasan Choudhury, Jian Wu, William A. Ingram, and Edward A. Fox. 2020. A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 515–516. https://doi.org/10.1145/3383583.3398590

[5] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[6] Jung-Ran Park. 2009. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly* 47, 3-4 (2009), 213–228. https://doi.org/10.1080/01639370902737240

[7] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[8] S. Rohatgi, C. Giles, and J. Wu. 2021. What Were People Searching For? A Query Log Analysis of An Academic Search Engine. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, Los Alamitos, CA, USA, 342–343. https://doi.org/10.1109/JCDL52503.2021.00062

[9] Alice Tani, Leonardo Candela, and Donatella Castelli. 2013. Dealing with metadata quality: The legacy of digital library efforts. *Inf. Process. Manag.* 49, 6 (2013), 1194–1205. https://doi.org/10.1016/j.ipm.2013.05.003

[10] S. Uddin, B. Banerjee, J. Wu, W. A. Ingram, and E. A. Fox. 2021. Building A Large Collection of Multi-domain Electronic Theses and Dissertations. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE Computer Society, Los Alamitos, CA, USA, 6043–6045. https://doi.org/10.1109/BigData52589.2021.9672058

[11] J. Wu, K. Williams, M. Khabsa, and C. Giles. 2014. The impact of user corrections on a crawl-based digital library: A CiteSeerX perspective. In *2014 International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*. IEEE Computer Society, Los Alamitos, CA, USA, 171–176. https://doi.ieeecomputersociety.org/