The Link between Dissertation Metadata Completeness and User Engagement in an Institutional Repository

Behrooz Rasuli[1*], Michael Boock[2], Joachim Schöpfel[3], and Brenda Van Wyk[4]

[1] Information and Society Research Department, Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran (Corresponding Author)
[2] Oregon State University, Oregon, USA
[3] University of Lille, Lille, France
[4] The University of Pretoria, Pretoria, South Africa

Author Note

Behrooz Rasuli https://orcid.org/0000-0001-6091-6967. Email: rasuli@irandoc.ac.ir
Michael Boock https://orcid.org/0000-0003-4248-5109. Email: michael.boock@oregonstate.edu
Joachim Schöpfel https://orcid.org/0000-0002-4000-807X. Email: joachim.schopfel@uni-lille.fr
Brenda Van Wyk https://orcid.org/0000-0003-3898-7042. Email: brenda.vanwyk@up.ac.za

Correspondence concerning this article should be addressed to Dr. Behrooz Rasuli, Information and Society Research Department, Iranian Research Institute for Information Science and Technology (IranDoc). Email: rasuli@irandoc.ac.ir

Disclosures and declarations
Authors have no conflicts of interest to disclose.

Funding
Not Applicable.

Acknowledgment
We would like to thank the MIT Libraries for their invaluable assistance with our research, particularly in answering our questions about the process of depositing dissertations. Their support greatly enhanced our understanding of the subject. Additionally, we would like to acknowledge Dr. Amir Hossein Seddighi, an assistant professor at IranDoc, for his significant contributions to our data collection efforts. We also extend our gratitude to the participants of the 27th International Symposium on Electronic Theses and Dissertations (ETD 2024) for their engaging feedback and comments, which have enriched our work and provided new perspectives. Finally, we appreciate the support of our colleagues and peers who encouraged us throughout this research.

The Link between Dissertation Metadata Completeness and User Engagement

Abstract

This study investigates the role of metadata quality in Electronic Theses and Dissertations (ETDs), focusing on its completeness and its impact on discoverability and user engagement within institutional repositories (IRs). Using DSpace@MIT as a case study, the current research analyzed 22,276 doctoral dissertations to assess metadata completeness and its correlation with the number of views and downloads. Various metadata fields and usage statistics were extracted for detailed analysis. The study identified a moderate positive correlation between the numbers of unique metadata fields and both the Department Views Ratio (DVR) and Department Download Ratio (DDR), suggesting that enriched metadata can improve the visibility and accessibility of dissertations. Additionally, the length of abstracts is positively correlated with engagement metrics (significance level for all reported results <0.001). In contrast, title length does not significantly influence the visibility. These findings showed the importance of high-quality metadata in enhancing the discoverability of ETDs. Three limitations are discussed, including the focus on a single repository, the lack of control for other variables that may impact user engagement, and the massive upload of thousands of theses in 2005. This research not only emphasizes the necessity of high-quality metadata for enhancing discoverability but also positions it as a strategic asset that can significantly amplify the visibility and impact of scholarly work. As institutions strive to foster open access and maximize research dissemination, our study provides actionable insights that can guide repository managers in refining their metadata practices.

Keywords: Visibility, Bibliometrics, Open Repository, Open Science, Altmetrics, Research Impact.

Introduction

Over the past three decades, higher education institutions (HEIs) have increasingly adopted digital formats for theses and dissertations (TDs) to enhance accessibility, visibility, and impact. These Electronic Theses and Dissertations (ETDs) are now widely discoverable through various channels, including institutional ETD program portals (e.g. Electronic Theses & Dissertations at Johns Hopkins University), national ETD portals (e.g. ShodhGanga in India), regional ETD portals (e.g. DART-Europe E-theses Portal), and Institutional Repositories (e.g. DSpace@MIT). However, regardless of the access point, the quality of an ETD's metadata is crucial for several reasons, including discoverability, interoperability, assessment, and preservation.

Research Problem and Motivation

Repositories use various strategies to ensure high-quality metadata for ETDs, from requiring detailed input during deposit to policy-driven approaches (Kasonde & Phiri, 2023) and automated quality enhancements (Choudhury et al., 2023). While metadata quality lacks a universal definition, it is commonly evaluated based on different criteria, with accuracy, completeness, and consistency being the most emphasized (Park, 2009).

Despite extensive research on ETD metadata quality and its recognized importance, a gap exists in empirical studies on the impact of metadata completeness on ETD impact. This study aims to bridge this gap by investigating the relationship between ETD metadata completeness and the number of views/downloads in institutional repositories (IRs). The underlying assumption is that more complete metadata enhances ETD discoverability, leading to a potential increase in the number of views and subsequently the number of downloads.

This research will utilize dissertations archived on DSpace@MIT as a case study. Established in the early 2000s, DSpace@MIT is the institutional repository of the Massachusetts Institute of Technology (MIT) and houses scholarly works produced by its affiliated researchers. In addition to providing metadata for ETDs, DSpace@MIT leverages IRUS (Institutional Repository Usage Statistics) to track and report the number of views and downloads for each item within the repository (Roosa, 2024). As of April 26, 2024, DSpace@MIT has 22,353 doctoral dissertations in 30 distinct collections[1].

Purpose of the Present Study

The purpose of the present study is to investigate several key aspects of ETDs within the DSpace@MIT repository. Specifically, this research aims to determine the completeness of metadata for doctoral dissertations, identify the number of views and downloads of ETDs, and explore the relationships between metadata completeness and both the number of views and downloads. Additionally, the study will examine the correlation between the numbers of views and downloads of ETDs. Finally, the research seeks to identify which metadata fields within ETDs demonstrate consistently higher completeness rates compared to others.

Literature review

Background history of ETDs

ETD was first introduced in the early 1990s to facilitate digital access to students' theses and dissertations (Rasuli et al., 2019). HEIs have increasingly adopted these digital formats to enhance accessibility, visibility, and impact of knowledge assets and scholarly communication availed in open access (Adam & Kaur, 2021; Van Wyk & Mostert, 2014). Its availability and accessibility are by now a

[1] https://dspace.mit.edu/handle/1721.1/131022

well-entrenched practices within research repositories collections and services. Notable examples are institutional ETD program portals such as the ETDs at Johns Hopkins University; the national ETD portals of ShodhGanga in India; regional ETD portals such as the DART-Europe E-theses Portal. Additionally, many institutional repositories exist, with the Massachusetts Institute of Technology's DSpace@MIT serving as a prominent example of a high-functioning system.

Since 2000, several metadata standards have been developed (Glogoff & Forger, 2000) providing comprehensive sets of descriptive metadata elements essential for database creation. When used with web search tools, these standards offer the granularity necessary for effective resource discovery, while also being grounded in the principle of openness to enhance global e-research networks. This openness relies on interoperability, enabling systems to communicate with each other and exchange information in a usable format (Adam & Kaur, 2021).

Much of the previous research on ETDs has concentrated on implementation projects, success factors, and sustainability at a macro level. However, on a more micro level, research examining content through metadata as a measure of success has not been fully explored. The quality of an ETD's metadata is crucial for discoverability, interoperability, assessment, and preservation. Interest in further research on ETDs is growing, but Choudhury (2023) points out significant gaps regarding appropriate methodologies and frameworks for extracting information from ETDs. He specifically notes the lack of information on ETD segmentation, metadata extraction, metadata quality improvement, and parsing reference strings (Choudhury, 2023). The literature underscores the urgent need for further research into metadata quality, particularly the completeness of metadata as a key indicator of overall quality.

Metadata quality in research and practice is often assessed through completeness, accuracy, consistency, accessibility, conformance, provenance, and timeliness (Kumar et al., 2024). Notably, accuracy, completeness, and consistency are the most emphasized criteria in the literature (Park, 2009). Additionally, Kasonde and Phiri (2023) emphasize the paramount importance of complete metadata for ETDs within IRs.

The nature of metadata

A generally used explanation of metadata is still viewing it as data about data. However, this over-simplified description is widely criticized (see Comber et al., 2006) within metadata research. This description falls short when looking at its application, required standards, and description of sources using metadata according to metadata standards in ETDs. As yet, there is no universally agreed-upon definition describing metadata quality. When defining metadata its purpose should be clarified and emphasized. In this sense, authors (Alemneh, 2008; Tarver et al., 2014) relate that the purpose of quality and comprehensive metadata is to provide adequate, correct, and relevant information to enhance visibility and discoverability.

Within ETDs, data must be structured systematically and ideally ordered according to a recognized metadata schema, allowing for a range of record or object formats. Alemneh and Phillips (2016) remind that metadata in ETDs should not be seen generically. Kasonde and Phiri (2023) state that the purpose of well-defined metadata is to link the creators of the information object with its users in the most effective way. In managing and maintaining metadata interoperability, authentication, and security can be achieved. One must bear in mind that that the nature of information objects may vary from discipline to discipline. It also serves as a preservation measure for future accessibility.

Types of Metadata in Dissertations

In order to get clarity on what is meant by the completeness of metadata it is essential to first establish the different types of metadata about ETDs. Essentially, the ETD metadata is needed to support the full range of digital preservation activities and types. Alemneh et al. (2014) refers to four types, while others name six and even eight types of metadata. Initially, the following types of metadata were listed:

- Provenance metadata: records the origin or provides an historical context or source provenance, such as specifying the analog source material for a digital derivative.

- Structural metadata: captures physical structural relationships, such as which image is embedded within which file, as well as logical structural relationships, such as page order, in born-digital or digitized TDs.
- Technical metadata: captures format-specific technical information that applies to any file type, including information about the software and hardware on which the digital object can be rendered or executed, as well as checksums and digital signatures to ensure fixity and authenticity.
- Administrative metadata: provides provenance information regarding who has cared for the digital object and what preservation actions have been performed on it. It also provides rights and permission information that specifies embargoes and access of ETDs and which preservation actions are permissible (Alemneh et al., 2014).

Notably, some authors refer to six or eight different types of metadata in ETDs. They add statistical metadata, also known as process metadata, and reference meta data, as well as preservation and rights metadata. These types of metadata serve different roles in the management, discovery, and preservation of digital and physical resources.

Metadata standards

Metadata standards play a crucial role in ensuring that metadata is comprehensive and consistent, which directly impacts its completeness. There are various types of metadata standards, including Dublin Core (DC), AACR2, EAD, TEI, METS, MODS, LOM, AGLS, ONIX, Darwin Core, CDWA LITE, CIDOC CRM, ETD-MS, IPTC Core Schema for XMP, MARC 21 XML, NISO Metadata for Images in XML, Multimedia Content Description Interface, GILS, GEM, and DDI. These standards can be categorized into several groups, such as descriptive, administrative (which includes technical, digital provenance, and rights/access), preservation, structural, and meta-metadata. Each metadata standard addresses aspects such as the rights holder authority, access conditions (whether paid or free), reproduction conditions, language of the document, and the physical characteristics of the document (Anil Hirwade, 2011).

Each metadata standard comprises its own specific elements. In their study, Park and Richard (2011) conducted an assessment of the metadata element sets utilized in ETDs within Canadian academic repositories. They analyzed the formats and usage patterns of metadata elements across ten institutional repositories, categorizing the variations based on different types. Their findings revealed a notable degree of inconsistency and variation in the current metadata elements. Bruce and Hillmann (2004) defined seven characteristics that should be considered when assessing the quality of metadata: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. It stands to reason that these characteristics must be evaluated for effectiveness. But, Stvilia et al. (2007) described 32 parameters in order to evaluate metadata quality. They also defined completeness as an important parameter for assessing metadata quality. A metadata instance is complete if it includes all the essential details to accurately represent the resource it describes (Ochoa & Duval, 2009).

Metadata standards and the completeness of metadata are deeply connected. By providing a structured framework with defined elements, metadata standards ensure that all necessary information is captured, leading to more complete metadata records. They promote consistency, support interoperability, and enable the use of automated tools, all of which contribute to the overall completeness and quality of metadata in research repositories. Despite extensive research on ETD metadata quality and its recognized importance, a gap exists in empirical studies on the impact of metadata completeness on ETD impact.

Metadata discoverability and visibility

Studies over the past two decades alluded to problems that may impact the quality and completeness of metadata (Bruce & Hillmann, 2004; Kasonde & Phiri, 2023). An earlier study (Bruce & Hillmann, 2004) identified reasons for poor metadata quality. The study raised concerns about the

effective training of staff producing metadata and warned that incompleteness in records will impact interoperability and discoverability. Subsequent studies further stressed the importance of metadata quality and more specifically metadata completeness towards enhanced discoverability (see Delgado-Quirós & Ortega, 2024; Osman et al., 2023; Park & Richard, 2011).

In 2016 a seminal paper was published in the journal Scientific Data offering guidelines for the FAIR principles (Wilkinson et al., 2016). The guidelines provide criteria for better findability, accessibility, interoperability, and reusability for data publication and stewardship as core to enabling transparent research. In terms of metadata completeness, the principle of reusability is important where reusability requires detailed description where metadata that is richly described with a plurality of accurate and relevant attributes. It is furthermore important to meet domain-relevant community standards. Authors (Musen et al., 2022) warn that ensuring that metadata are richly described with a plurality of accurate and relevant attributes may posit challenges, as it is not at all times clear what exactly is considered as accurate. They also raised questions question about monitoring the implementation and adherence of the FAIR principles.

Metadata Quality

The quality of metadata in research repositories is crucial for ensuring that research outputs are easily discoverable, accessible, and usable. Metadata quality is typically evaluated based on its correctness, completeness, accuracy, consistency, clarity, relevance, interoperability, flexibility, and redundancy. High-quality metadata ensures that research outputs are properly indexed, easily found by others, and can be accurately cited, leading to broader dissemination and greater impact (Choudhury et al., 2023).

Both correctness and completeness of metadata are important in looking at quality of metadata. The correctness of metadata refers to the intellectual distance separating them from the true representation of the resource being described (Margaritopoulos et al., 2008). They explain correctness on two levels:

- The first, lower level concerns the requirement that the values of the metadata fields must obey the grammatical and syntactical rules of the language and the metadata standard or the application profile used. Missing letters, misspelled words, inconsistent formatting or representation of the same fields, and fields containing inappropriate values according to the standard, are among the problems of this level.
- The second, higher level of correctness requires the semantical rightness of the values of the metadata fields, that is, the true representation of reality and the absence of any deception (Margaritopoulos et al., 2008).

A metadata record must strictly follow the rules and guidelines of the standard or the application profile to be correct.

Factors impeding metadata quality

Much research describes factors that negatively impact the quality of TD (Chisale & Phiri, 2023; Kasonde & Phiri, 2023; Osman et al., 2023; Park & Richard, 2011). Open access harvesters such as OAI-PMH are crucial protocols for building connected and interoperable digital information infrastructure, and their usefulness largely depends on the quality and consistency of the metadata provided by open access repositories. Metadata harvesting by aggregation services such as the Networked Digital Library of Theses and Dissertations (NDLTD)'s Union Catalog and the Open Access Theses and Dissertations portal struggle to harvest poor quality metadata (Chisale & Phiri, 2023).

A lack of training of those responsible for assigning metadata as well as insufficient workflow monitoring and evaluation may lead to human error in capturing descriptive metadata (Phiri, 2020). Linked to this could be inadequate resource allocation. Looking at the differences in objects being described in ETDs; heterogeneity is another factor.

Metadata completeness as a subset of quality

In addition to the quality of metadata, its completeness within the definitions of metadata standards are critical determinant of the discoverability, visibility, and accessibility of particularly ETDs. Metadata quality in research and practice is often assessed through completeness, accuracy, consistency, accessibility, conformance, provenance, and timeliness (Kumar et al., 2024). Completeness of metadata refers to their sufficiency to fully describe a resource covering all its possible aspects. The completeness of metadata in TDs impacts the ability to retrieve reliable descriptive data by web crawlers, and the successful access the web the loss of information derived from integrating different sources.

Repositories managers use various methods to ensure ETDs are described in depth, and that the quality of metadata is high. These may include requesting researchers to provide more comprehensive information during ETD submissions, to having procedures and policies in place to evaluate and augment metadata (Kasonde & Phiri, 2023). There are further developments towards automated quality improvement (Choudhury et al., 2023).

Margaritopoulos et al. (2008) see the completeness of metadata as a subset of metadata quality. All necessary metadata fields should be filled in. This includes titles, authors, abstracts, keywords, dates, funder information, and any other relevant data. Incomplete metadata can hinder the discoverability and proper citation of the research. A paucity of literature on the completeness of metadata records in ETDs is evident and must be explored further.

Improving metadata quality and completeness

Poor quality metadata can be prevented if quality checks for completeness are in place. Both checking for correctness and completeness of metadata can be improved by adhering to metadata standards. Continuous training and updating of guidelines may elevate human error in capturing metadata. Further collaboration with the research community may enable refined and better-described metadata coming from diverse disciplines. Over and above these measures more and more calls are heard for automated tools in metadata generation (Chisale & Phiri, 2023). Improving metadata quality involves a combination of standardization, accuracy, richness, and interoperability, along with ongoing efforts to address challenges and adapt to new developments in the research landscape. Metadata quality analysis must be part of best practices and service standards.

Metadata policies

Metadata policies and institutional repositories are closely connected, as metadata is essential for organizing, accessing, and preserving digital assets over the long term. These policies define how metadata is created, managed, and used within repositories. Their main goals include ensuring consistency in metadata formats, enabling smooth data exchange between repositories, search engines, and indexing services, and improving search and retrieval for users.

Research on metadata policies in institutional repositories shows significant variation, partly due to diverse metadata sources and structures (Chapman et al., 2009). However, most policies rely on established metadata standards, such as METS and MODS from the Library of Congress, ETD-MS for electronic theses and dissertations, and, most commonly, the simple and widely accepted Dublin Core (Park et al., 2015).

Recently, metadata policies have increasingly aligned with the FAIR principles, aiming to make repository content more findable, accessible, interoperable, and reusable, and have been assessed accordingly to measure their degree of FAIRness (d'Aquin et al., 2023). One key challenge is transparency—many institutional repositories do not clearly document their metadata policies, making it difficult to assess their approaches (Kenfield, 2019).

ETDs at MIT

This study investigates the relationship between ETD metadata completeness and its impact on discoverability within the institutional repository of the Massachusetts Institute of Technology (MIT). As

a private land-grant research university MIT was established in 1861 in Massachusetts, USA. MIT was one of the key institutions responsible for the creation of the DSpace institutional repository software platform. Developed in collaboration with Hewlett-Packard, it was launched in 2002 to facilitate the management and dissemination of digital content, including theses and dissertations (Baudoin & Branschofsky, 2003). The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) was adopted, allowing DSpace to make its Dublin Core-formatted metadata available to compatible harvesting code. DSpace@MIT contains more than 53,000 selected theses and dissertations from all MIT departments, although it does not include all MIT theses. MIT Libraries encourage students to incorporate appropriate accessibility features and metadata into their thesis documents (MIT Libraries, 2024a, 2024b).

MIT Libraries staff are responsible for ensuring the accuracy and completeness of the final metadata for all thesis records. Since 2022, MIT has encouraged authors to submit their own metadata to the libraries when they deposit their theses, although this submission is not mandatory. Even if authors do not provide metadata, the libraries staff review and update the records, adding any missing information as necessary to maintain the quality and discoverability of repository theses and dissertations.

Method

Data Collection

This study collected data on 22,276 doctoral theses from the DSpace@MIT repository. The data collection process involved extracting the metadata fields and usage statistics for each dissertation. MIT was chosen as a case study for several reasons:

- **High Visibility**: As a renowned university with international reach, MIT dissertations likely attract a significant number of views and downloads, providing a robust dataset for analysis.
- **Controlled Variables**: Focusing on DSpace@MIT as a single institutional repository was a deliberate choice to control for the external influence of repository reputation on user engagement metrics, thereby potentially mitigating its impact on the findings. However, it is important to acknowledge that while this approach helps to limit the variability introduced by differing institutional reputations, other external factors beyond the scope of this study may still influence user behavior.
- **Standardized Format**: Limiting the study to doctoral dissertations ensures a consistent type of work across the sample, minimizing the impact of document type as a confounding factor.
- **Data Availability**: DSpace@MIT offers both comprehensive metadata for dissertations and readily accessible usage statistics for each document, facilitating metadata completeness and discoverability data collection.

The following steps outline the data collection methodology:

1. **Data Source**: The primary source of data was the DSpace@MIT repository, specifically the URL format: https://dspace.mit.edu/handle/1721.1/XXXXX?show=full, where XXXXX represents the unique ID for each dissertation. To extract the IDs, the Doctoral Theses collection in DSpace@MIT (https://dspace.mit.edu/handle/1721.1/131022/recent-submissions) was browsed, and all submitted records until August 21, 2024, were retrieved.
2. **Metadata Fields**: The study focused on metadata fields that begin with "dc." (Dublin Core), which are standard fields used for describing digital resources.
3. **Data Extraction**: The data extraction process was conducted from August 10 to August 21, 2024. Each dissertation's metadata and usage statistics were retrieved and recorded.

Data Organization

Once the data was collected, it was organized into an Excel file with the following columns for each dissertation:

1. **Record ID**: A unique identifier assigned by DSpace@MIT to each record.
2. **Department**: This refers to the academic department at MIT associated with the dissertation. The original dataset included more than 70 different values for this field. However, some dissertations

contained misspellings or variations in the names of specific departments. To address this issue, all department and program names were standardized by consulting the MIT website, resulting in a cleaned and unified dataset.

3. **Date Available**: The date the dissertation became publicly accessible through DSpace@MIT. 2005 is the first year for the availability of dissertations in the current study's dataset. For dissertations lacking a specific "Date Available," the study utilized the "Date Issued" or "Date Copyright" fields as substitutes to ensure completeness in the dataset.
4. **Number of Unique Metadata Fields**: This refers to the count of distinct metadata properties that are filled out, excluding duplicates. For instance, if a specific field (e.g., dc.description.abstract) is filled out multiple times for one record, it is counted as a single field in the overall numeration. For example, this record at https://dspace.mit.edu/handle/1721.1/147290?show=full has 17 unique metadata fields.
5. **Number of Duplicated Metadata Fields**: This represents the total count of all metadata fields filled out, including duplicates. For example, some records may include multiple entries for certain fields, such as "dc.contributor.advisor" when a dissertation has co-supervisors. For example, this record at https://dspace.mit.edu/handle/1721.1/147290?show=full has 19 duplicated metadata fields.
6. **Abstract Word Count**: Total word count of the dissertation's abstract, considering all fields where the abstract may be recorded.
7. **Title Word Count**: Total word count of the dissertation's title.
8. **Number of Downloads**: Total full-text downloads recorded for the dissertation.
9. **Department Download Ratio (DDR)**: DDR is a download-based measure of the impact of one record, and it indicates the relative download performance of a dissertation when compared to similarly aged dissertations in its department. It is calculated by dividing the number of downloads by the geometric mean of downloads for similarly aged dissertations in the same department. This indicator ensures the normalization of the number of downloads across different departments and years.
10. **Number of Views**: Total page views recorded for the dissertation.
11. **Department Views Ratio (DVR)**: DVR is a view-based measure of the visibility of one record, and it indicates the relative page view performance of a dissertation when compared to similarly aged dissertations in its department. It is calculated by dividing the number of views by the geometric mean of views for similarly aged dissertations in the same department. This indicator ensures normalization of the number of page views across different departments and years.
12. **PDF Size**: The size of the attached digital file in Megabytes.

Data Analysis

The analysis focused on several key aspects to understand the relationship between metadata completeness and user engagement. First, the study assessed metadata completeness by analyzing the number of unique and duplicated metadata fields for each record across the dataset and through descriptive statistics. Next, the analysis examined the correlation between metadata completeness and usage statistics, specifically the number of page views and downloads. Statistical tests were conducted to determine whether there was a significant relationship between the completeness of metadata and these engagement metrics. This step was crucial for understanding how metadata quality might influence user interaction with the dissertations.

Results

As stated earlier, this study collected data on 22,276 doctoral theses from the DSpace@MIT repository. Figure 1 and

Figure ***2*** illustrate the frequency of dissertations across various departments and years. Notably, the Department of Electrical Engineering and Computer Science stands out with a total of 3,331 dissertations. Additionally, the year 2005 was significant, as it saw the release of 5,578 dissertations through DSpace@MIT, partly with issue dates earlier than 2005 but made available only in 2005.

Figure 1

Distribution of Dissertations by Year at DSpace@MIT

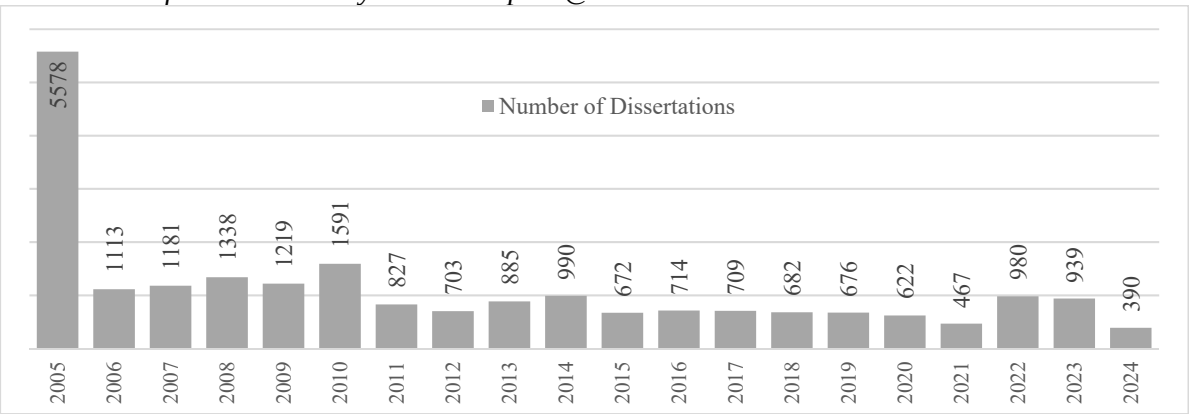
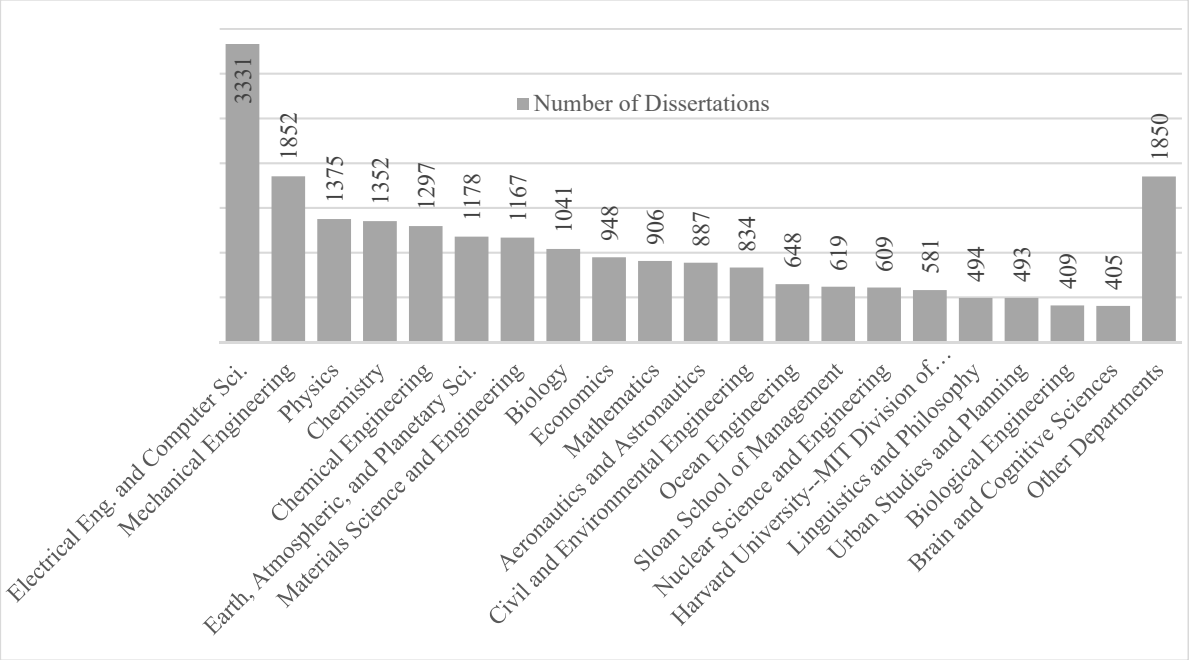


Figure 2

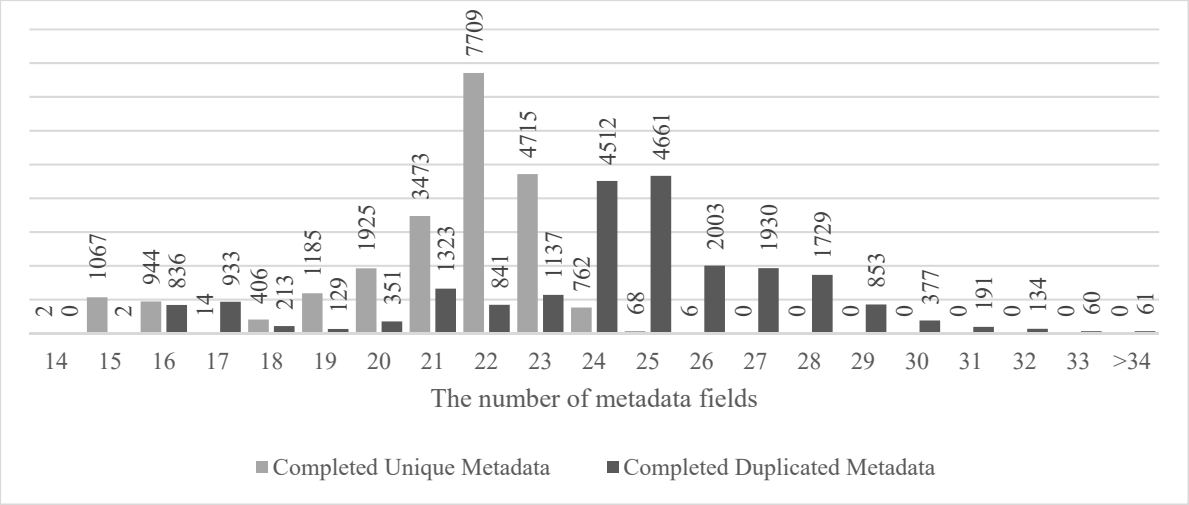
Distribution of Dissertations by Department at DSpace@MIT



The analysis of unique and duplicated metadata fields completed for the doctoral dissertations in the DSpace@MIT repository reveals significant insights into the quality of metadata associated with these academic works. Figure 3 presents the distribution of completed unique and duplicated metadata fields across the dissertations.

Figure 3

Distribution of Completed Unique and Duplicated Metadata Fields in Doctoral Dissertations at DSpace@MIT

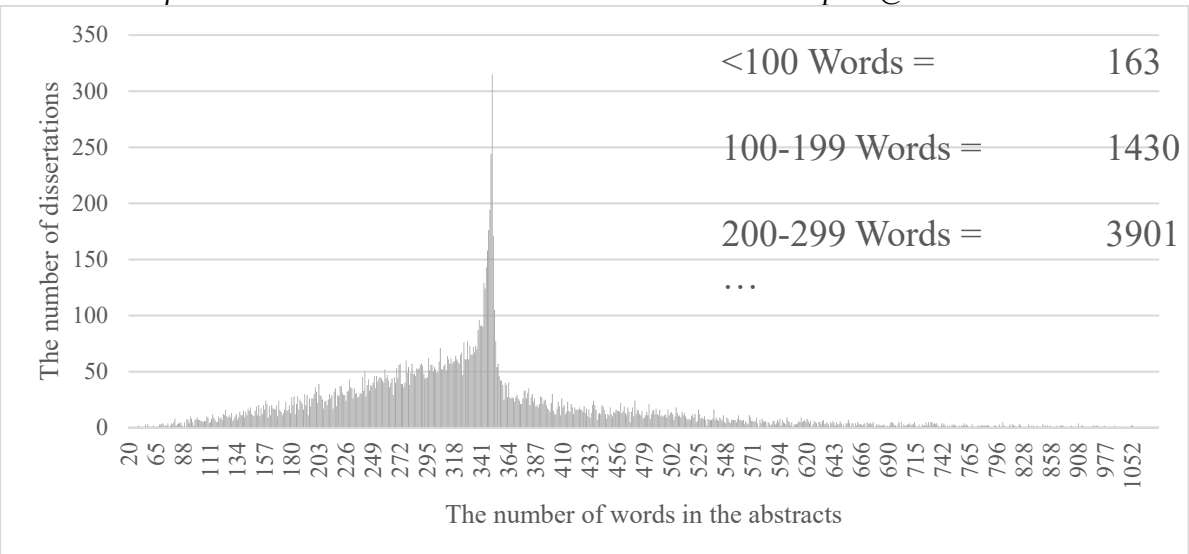


According to Figure 3, the minimum number of unique metadata fields completed is 14, while the maximum is 26. However, the majority of dissertations, totaling 7,709, have 22 unique metadata fields completed. On the other hand, the number of duplicated metadata fields varies significantly, with a minimum of 15 fields and a maximum of 41 fields completed. Notably, the highest frequency of duplicated fields occurs at 24 fields, with 4,512 instances recorded. This suggests that many dissertations include multiple entries for certain metadata categories, which may enhance the detail and context of the information provided. However, the assumption that more entries indicate greater completeness should be interpreted cautiously, as multiple entries do not necessarily improve metadata quality. For instance, while fields such as "contributor.advisor" may contain multiple entries when a dissertation has co-supervisors, fields like "contributor.author" are generally expected to have only a single entry.

Abstracts serve as critical components of academic work, offering a concise overview of the research objectives, methods, and findings. Understanding the distribution of word counts can help assess the variability in how researchers communicate their work. Figure 4 presents the distribution of word counts for abstracts across the doctoral dissertations in the DSpace@MIT repository.

Figure 4

Distribution of Abstract Word Counts in Doctoral Dissertations at DSpace@MIT



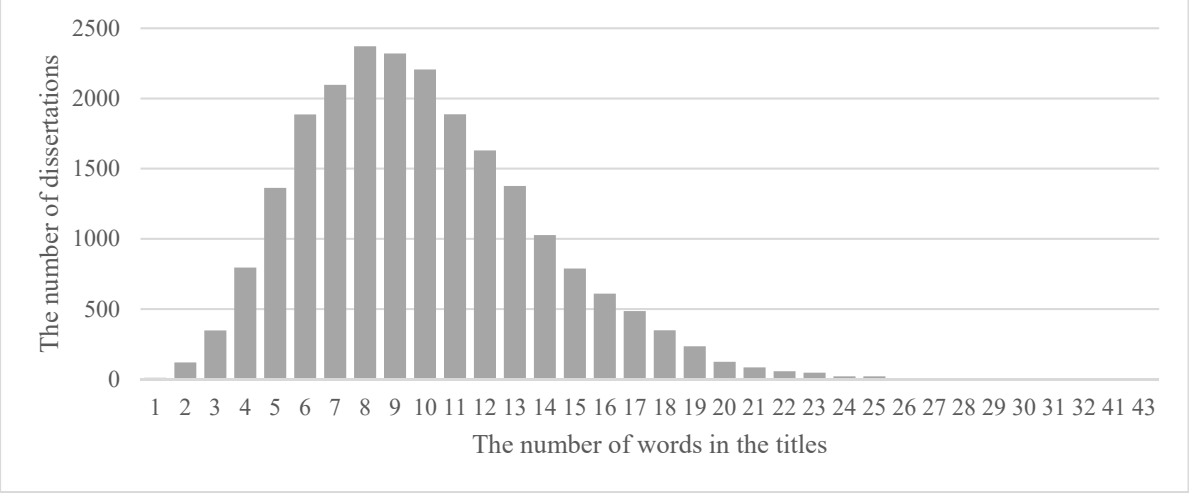
According to Figure 4, after analyzing the abstract of 14,343 dissertations (didn't have an abstract), the data reveals a wide range of abstract lengths, with some dissertations without and abstract (7,933 records) and some exceeding 1,000 words. However, the majority of abstracts fall within a more

moderate range, with the highest frequency observed in the 300-399 words category, which contains 5908 records. This suggests that researchers tend to favor a length that allows for a comprehensive yet concise presentation of their work.

In addition to abstracts, the titles of dissertations also play an important role in the visibility and impact of academic publications. Figure 5 presents the distribution of word counts for titles across the doctoral dissertations in the DSpace@MIT repository.

Figure 5

Distribution of Title Word Counts in Doctoral Dissertations at DSpace@MIT



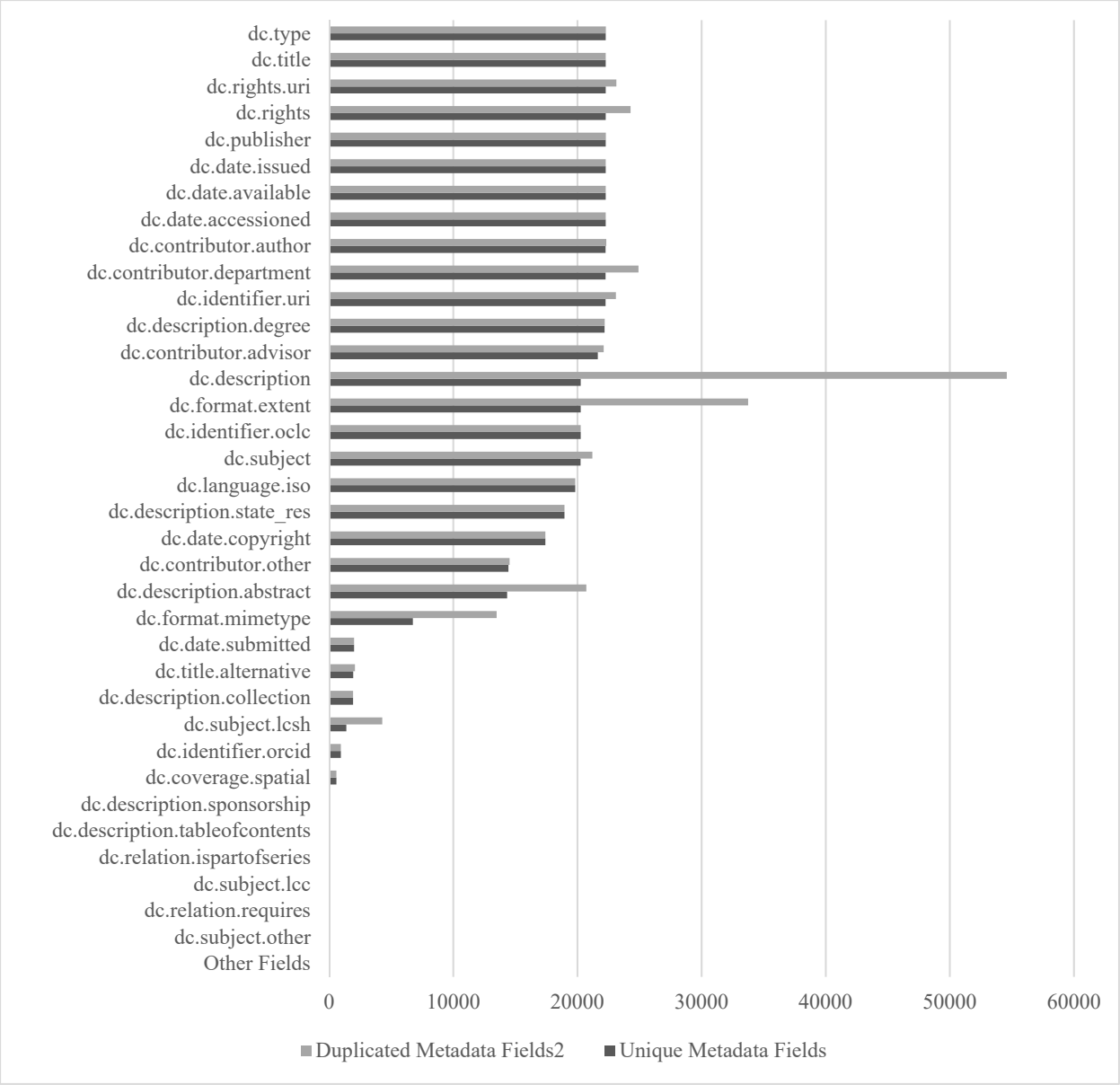
According to Figure 5, the distribution reveals a wide range of title lengths, with the number of titles varying significantly across different word counts. The most frequently occurring title length is 8 words, with a total of 2,371 titles recorded. Following closely, the 7-word titles also show a high frequency, with 2,096 instances, and the 9-word titles have 2,320 titles.

Another important aspect of examining metadata completeness is identifying which metadata fields are filled out most frequently. Figure 6 presents the results of the analysis regarding the frequency of various metadata fields that are completed for the dissertations in DSpace@MIT.

Figure 6 illustrates the frequency of unique and duplicated metadata fields for doctoral dissertations in the DSpace@MIT repository. Each entry represents a specific metadata field, displaying the count of unique entries alongside the total number of duplicated entries for that field. For example, the "dc.description.abstract" field shows a significantly high count of 14,343 unique entries and 20,714 duplicated entries (it means that there are 6,371 records with two or more than two fields for "dc.description.abstract"), indicating its frequent use. In contrast, fields such as "dc.relation.requires" and "dc.subject.other" exhibit lower counts, suggesting they are less commonly filled out. The other fields (e.g. dc.contributor, dc.date, dc.date.created, dc.date.updated, dc.identifier.govdoc, dc.relation, dc.language, dc.audience.educationlevel, and dc.identifier.other) are filled out for only 27 dissertations.

Figure 6

Frequency of Unique and Duplicated Metadata Field Types for Doctoral Dissertations in DSpace@MIT



The interpretation of Figure 6 requires further clarification. This figure presents the frequency of unique and duplicated metadata fields for doctoral dissertations in the DSpace@MIT repository. Each entry represents a specific metadata field, displaying the count of unique entries alongside the total number of duplicated entries for that field. For example, the "dc.description.abstract" field shows a significantly high count of 14,343 unique entries and 20,714 duplicated entries, indicating that 6,371 records contain multiple entries in this field. However, certain fields, such as "dc.subject.lcc," have no values at all. This raises questions about whether these fields are consistently left empty or simply underutilized in the metadata entry process.

The visibility and impact of academic research can be significantly influenced by the number of downloads and page views that dissertations receive. Figure 7 and Figure 8 present the distribution of download and page view statistics for doctoral dissertations across various MIT departments. The values for the median (MD), geometric mean (GM), and total number of downloads/page views per department are shown in the figures.

Figure 7

Download Statistics by Department for Doctoral Dissertations at DSpace@MIT (log scale)

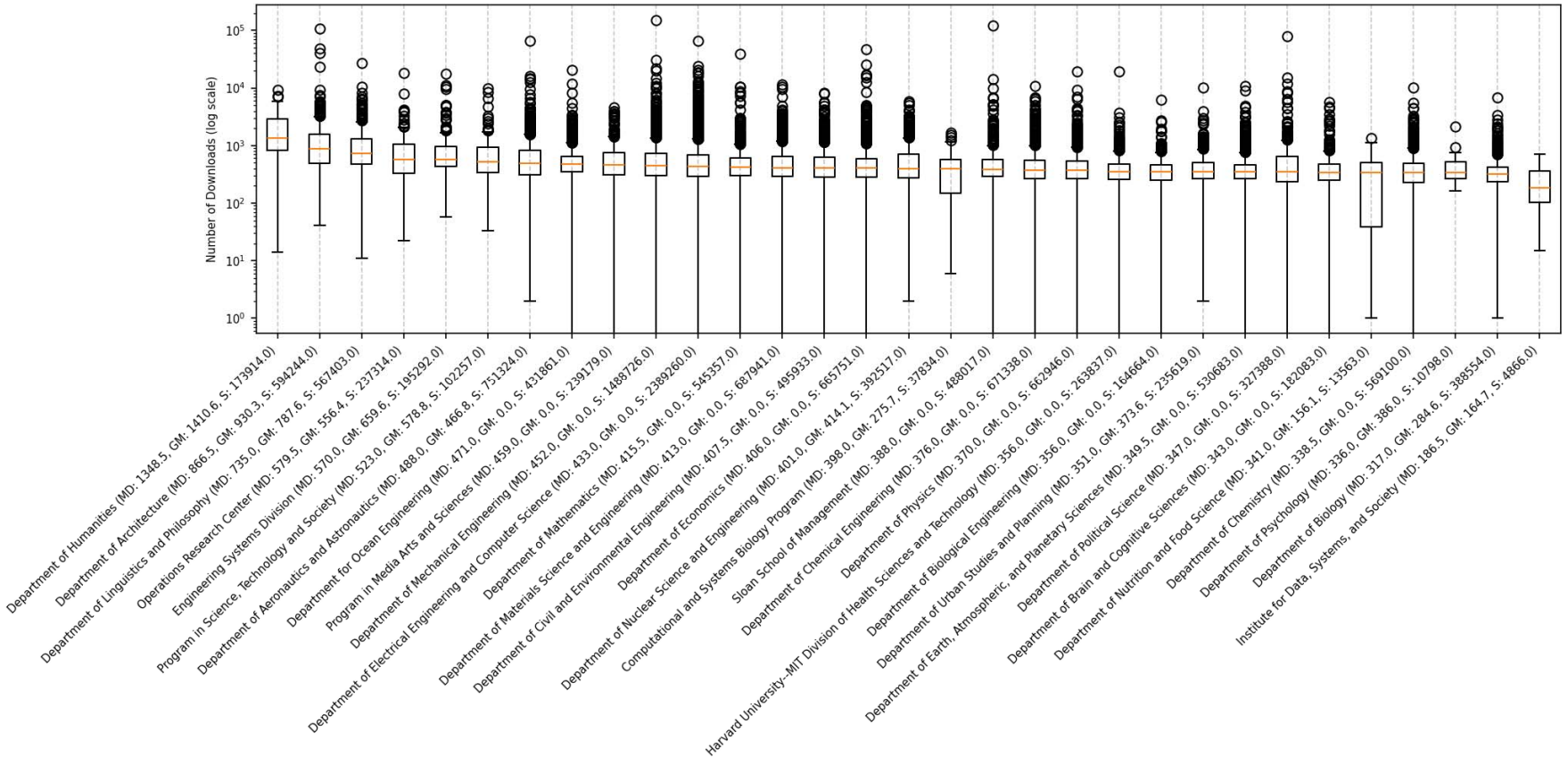
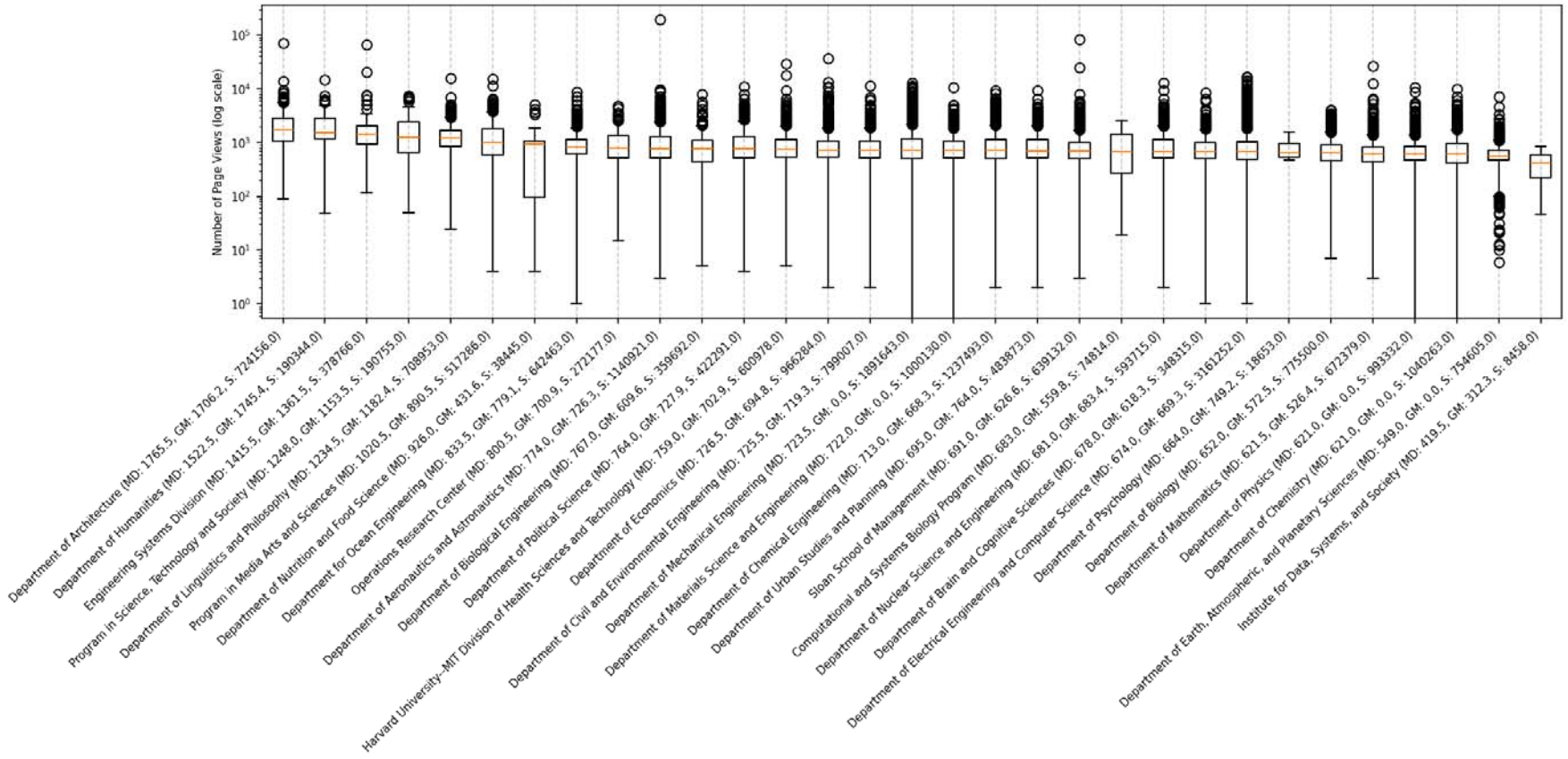


Figure 8

Page View Statistics by Department for Doctoral Dissertations at DSpace@MIT (log scale)



According to Figure 7, the box plot demonstrates significant variation in download statistics between departments. The Department of Humanities has the highest median download count at 1,348.5, followed closely by the Department of Architecture with 866.5 downloads. In contrast, departments such as the Institute for Data, Systems, and Society report far lower download statistics, with a median of 186.5 downloads. The total number of downloads is notably high in the Department of Electrical Engineering and Computer Science, with 2,389,260 downloads.[2]

According to Figure 8, The box plot reveals a diverse range of page view counts among departments. The Department of Architecture exhibits the highest median page views at 1,765.5, followed by the Department of Humanities at 1,522.5 views. On the lower end, departments such as the Institute for Data, Systems, and Society have a median of 419.5 views. The Department of Electrical Engineering and Computer Science leads in total page views, with a sum of 3,161,252.

In order to investigate the possible correlations between dissertation metadata quality and user engagement, statistical analyses are needed. As a first step, examining descriptive statistics provides a foundational understanding of the distribution and variability of key variables across the dataset. Table 1 provides a detailed snapshot of the dissertations' metadata quality and user engagement levels within the DSpace@MIT repository.

Table 1

Descriptive Statistics of Dissertation Metadata and Access Metrics in DSpace@MIT

Variables	Descriptive Statistics								
	N	*Min*	*Max*	*Sum*	*Mean*	*Median*	*SD*	*Variance*	*SK*
Downloads	22276	0	151810	14509563	651.35	403.0	2053.36	4216289.89	41.60
DDR	22276	0	301	31364	1.41	0.918	3.78	14.26	43.52
Page View	22276	0	195261	21646075	971.72	702.0	1868.49	3491265.40	59.96
DVR	22276	0	180	27153	1.22	0.949	1.89	3.55	52.43
Unique Metadata	22276	14	26	470818	21.14	22.00	2.19	4.80	-1.47
Duplicated Metadata	22276	15	41	542847	24.37	25.00	3.37	11.38	-.58
Abstract Word Count	22276	0	1973	4852788	217.85	251.0	195.76	38320.78	.61
Title Word Count	22276	1	43	220604	9.90	9.000	3.96	15.70	.73
PDF Size (MB)	22276	0.05	524	420854	18.89	13.60	20.72	429.22	6.42

According to Table 1, Downloads range from 0 to 151,810, with a mean of 651.35 and a median of 403.0, indicating a right-skewed distribution (skewness of 41.60) where a small number of dissertations have exceptionally high download counts. Similarly, page views vary from 0 to 195,261, with a mean of 971.72 and a median of 702.0, and an even higher skewness of 59.96. The number of unique metadata fields completed per dissertation ranges between 14 and 26, with an average of 21.14 and a median of 22.00. Document characteristics also exhibit notable findings. The average abstract word count is 217.85 words, with a median of 251 and a maximum of 1,973 words. Titles average

[2] The geometric mean of zero for some departments, such as the Department of Chemical Engineering, indicates that there is (at least) a dissertation with no downloads or page views. The geometric mean is calculated by multiplying all the values together and then taking the nth root (where n is the number of values). If any of the values in this calculation are zero, the result will also be zero.

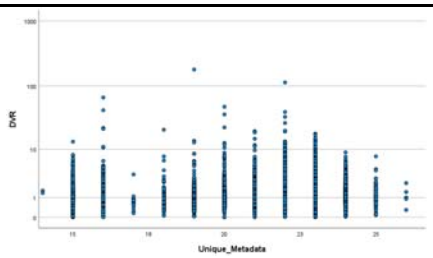
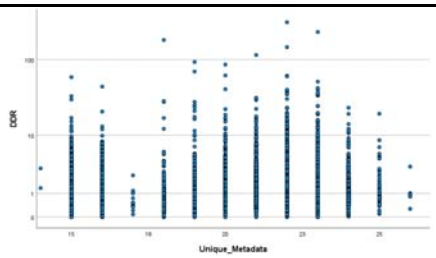
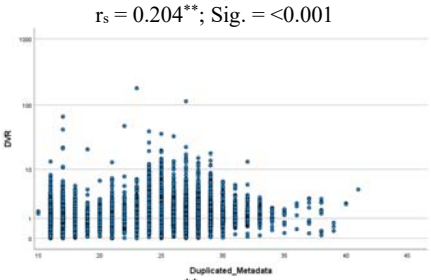
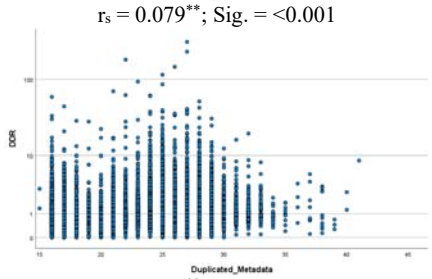
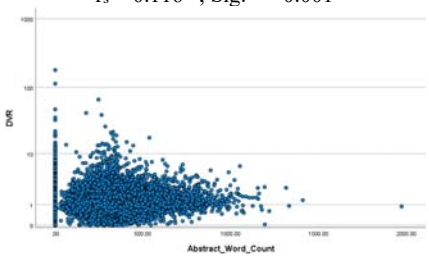
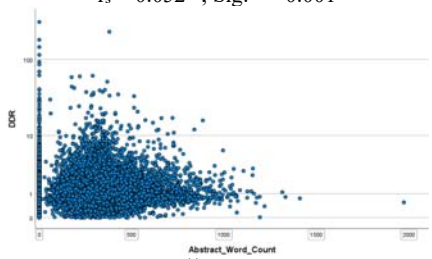
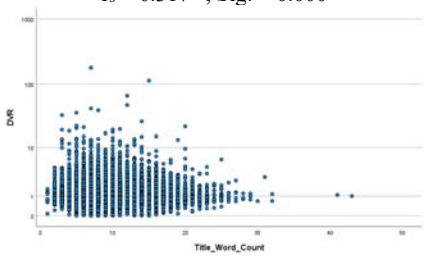
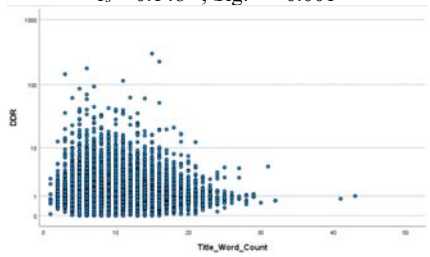
9.90 words in length, with a median of 9 and a maximum of 43 words. The PDFs have an average size of 18.89 MB.

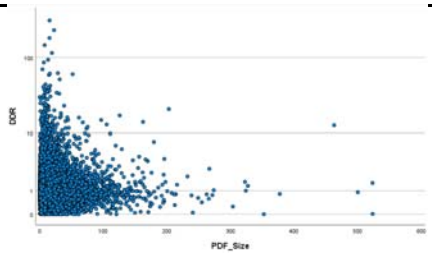
To further explore the relationships between dissertation metadata quality and user engagement, a Spearman rank correlation test was conducted. The results of the Spearman correlation analysis are summarized in Table 2. The table presents the correlation coefficients (r_s) along with their significance levels (p-values [Sig. (2-tailed)]).

The Department Download Ratio (DDR) and Department Views Ratio (DVR) are important metrics used to assess the performance of dissertations within their respective academic departments. Both ratios provide a normalized measure of engagement, allowing for fair comparisons among dissertations that may differ significantly in terms of their age, department, and overall visibility.

Table 2

Correlation Between Dissertation Metadata Quality/Characteristics and Engagement Metrics (log scale)

Variables	DVR (N = 22,276)	DDR (N = 22,276)
Number of Unique Metadata	 $r_s = 0.204^{**}$; Sig. = <0.001	 $r_s = 0.079^{**}$; Sig. = <0.001
Number of Duplicated Metadata	 $r_s = 0.116^{**}$; Sig. = <0.001	 $r_s = 0.052^{**}$; Sig. = <0.001
Number of Abstract's Word	 $r_s = 0.317^{**}$; Sig. = 0.000	 $r_s = 0.148^{**}$; Sig. = <0.001
Number of Title's Word	 $r_s = -0.009$; Sig. = 0.204	 $r_s = -0.048^{**}$; Sig. = <0.001

Variables	DVR (N = 22,276)	DDR (N = 22,276)
PDF Size	Not Applicable	 r_s = -0.122**; Sig. = <0.001

** *Correlation is significant at the 0.01 level (2-tailed).*

In Table 2, the number of unique metadata fields and the number of duplicated metadata fields are analyzed in relation to DVR and DDR. Unique metadata fields refer to distinct pieces of information describing a dissertation, such as title, author, abstract, and keywords. A higher number of unique metadata fields generally enhances discoverability by providing more structured information for indexing and retrieval. Duplicated metadata fields, on the other hand, occur when a same metadata field is duplicated, either due to inconsistencies in data entry or system design.

According to Table 2, there is a moderate positive correlation between the number of unique metadata fields and both DVR (rs = 0.204, Sig. = <0.001) and DDR (rs = 0.079, Sig. = <0.001). This indicates that dissertations with a greater number of unique metadata fields tend to have higher visibility (DVR) and download performance (DDR) compared to their peers. The stronger correlation with DVR suggests that unique metadata may be particularly effective in attracting views, which could lead to increased downloads.

Additionally, a positive correlation exists between the number of duplicated metadata fields and both DVR (r_s = 0.116, Sig. = <0.001) and DDR (r_s = 0.052, Sig. = <0.001), though the correlation is weaker than that of unique metadata. This suggests that while some duplication may contribute to engagement, its impact is relatively minor compared to unique metadata.

Discussion

Metadata serves as a fundamental pillar in our information-centric society, facilitating the efficient organization, retrieval, and utilization of information (Tani et al., 2013). To ensure ETDs are read and used, high-quality metadata is essential. The ETD community employs a variety of metadata practices at national, regional, and international levels. The criteria for what define minimal, good, and optimal metadata—regardless of format or schema—varies based on numerous factors, which can differ between countries and institutions. High-quality metadata is crucial for creating a union catalog, which is essential for library networking and resource sharing. Although ETDs are managed by the institutions that produce them, it is possible to present them as a unified collection through a central search engine that aggregates all the metadata. When users find a relevant document, they are redirected to the institution that holds it (Alemneh et al., 2014).

The findings revealed that out of the 129 metadata fields in the Dublin Core schema (DCMI Usage Board, 2020), 44 fields are utilized to describe MIT's dissertations. The average completeness of unique fields across the records is 21, while the average for duplicated fields is 24. These results indicate that the metadata for dissertations in DSpace@MIT is not fully complete, suggesting potential areas for improvement in metadata quality.

The moderate positive correlation between the number of unique metadata fields and both DVR and DDR indicates that more detailed metadata can (relatively) enhance the visibility and accessibility of dissertations. Completeness in metadata refers to the inclusion of all necessary information that accurately describes a resource (Ochoa & Duval, 2009). When metadata is comprehensive, it enhances the ability of users to find and access relevant materials more easily. Moreover, complete metadata not only facilitates easier discovery but also contributes to the credibility of the repository (Fear & Donaldson, 2012). When users are provided with sufficient information about a dissertation, they will have a clear understanding of the content of the described resource, potentially leading to more downloads (or helping users determine if the dissertation is

worth reading). According to Tani et al. (2013), the quality of metadata directly impacts the discoverability of information resources.

However, it is important to note that the correlation between metadata completeness and usage is moderate (and not strong), which may suggest that other factors also play a significant role in user engagement. One possible explanation for this is that ETDs may be crawled and retrieved based on their full text, which may diminish the relative importance of metadata in some contexts. This raises an important distinction between the relevance of FAIRness (Findable, Accessible, Interoperable, and Reusable) in terms of metadata completeness for machine readability and the human access facilitated by discovery tools, where the full text may take precedence.

The correlation between abstract length and engagement metrics suggests that more detailed abstracts may attract greater attention, as this factor can predict a higher score of DVR. The Abstract is one of the primary marketing elements of any scientific output (Pottier et al., 2023). When abstracts contain a rich array of pertinent terms and phrases, they become more likely to be indexed effectively by web search engines, union catalogs, and the DSpace@MIT repository's search engine. This increased visibility can lead to higher search rankings, making it easier for potential readers to find the dissertation. Moreover, the presence of relevant terminology not only aids in search engine optimization (SEO) but also ensures that the content aligns with the interests and queries of the target audience. By using specific and widely recognized terms related to their field, researchers can attract a more relevant readership, thereby increasing the likelihood of engagement with their work. Previous studies have found that the words in abstracts correlate with the citations of research outputs, with citation counts increasing steadily as abstract length increases (Robson & Mousquès, 2016; Sohrabi & Iraj, 2017).

Interestingly, the analysis reveals that title length does not significantly impact visibility, and longer titles may even correlate with lower download rates. This finding challenges the assumption that more descriptive titles necessarily lead to higher engagement, suggesting that brevity may be more effective in capturing reader interest. Supporting this finding, research by Paiva et al. (2012) found that articles with shorter titles had higher viewing and citation rates compared to those with longer titles.

Additionally, the negative correlation between PDF size and download rates indicates that users may prefer smaller, more manageable files, which could inform future guidelines for dissertation submissions. Users may be less inclined to download larger files possibly due to concerns about download time, storage space, or perceived accessibility.

Limitations, Recommendations, and Conclusion

Metadata completeness and accuracy play an important role in the retrieval of scientific publications (Céspedes et al., 2024). This study highlighted the role of metadata completeness in influencing user engagement within the DSpace@MIT repository. The findings showed the importance of maintaining high standards for metadata completeness, as this factor is closely linked to the visibility and accessibility of academic output.

The slightly negative correlation between title length and user engagement may surprise but confirms the results of other studies (such as Jamali & Nikzad, 2011) on the impact of title type and length on downloads and citations. While shorter titles tend to be correlated with higher views and downloads, the relationship is not absolute. The clarity, relevance, and appeal of the title seem more important than just its length. A balance between conciseness and informativeness is key.

One significant limitation is the focus on a single repository, which may not fully represent the broader landscape of academic repositories. The future research should consider a comparative analysis of multiple repositories. Additionally, the study primarily relied on two quantitative metrics, which may overlook qualitative aspects of user engagement, such as user satisfaction and the context in which dissertations are accessed. Incorporating qualitative methods, such as user surveys or interviews, could provide deeper insights into how metadata influences user behavior and engagement.

Another limitation was that we could not control for several indicators that may impact user engagement, such as the quality of the dissertations' content, the relevance of the dissertation topics to users' interests, and the demographics of the users accessing the repository. These factors can

significantly influence how often dissertations are viewed and downloaded. For instance, dissertations that address trending or highly relevant topics may naturally attract more attention, regardless of the completeness of their metadata. To gain a more comprehensive understanding of user engagement, future research can study these contextual factors.

The fact that 25% of the dissertations were uploaded in 2005 may have an impact on the statistical reliability of the correlation between usage statistics and metadata, because of age effect and time bias, and because of the overrepresentation of older dissertations. However, the choice of the DVR and DDR indicators contributes to limit this potential bias, as they don't compare the number of user engagement of 2005 dissertations to the 2020 dissertations but, instead, the statistics of a dissertation published in 2005 to all the others published in 2005 in the same field. However, a further statistical analysis was done after excluding 2005 uploaded records and the results showed a significant correlation between DVR and Number of Unique Metadata (r_s = 0.154 and Sig. = <0.001), DVR and Number of Duplicated Metadata (r_s = 0.124 and Sig. = <0.001), DVR and Number of Abstract's Word (r_s = 0.051 and Sig. = <0.001), DDR and Number of Unique Metadata (r_s = 0.056 and Sig. = <0.001), DDR and Number of Duplicated Metadata (r_s = 0.060 and Sig. = <0.001), DDR and Number of Title's Word (r_s = -0.051 and Sig. = <0.001), and DDR and PDF Size (r_s = -0.139 and Sig. = <0.001). But the correlations between DVR and Number of Title's Word (r_s = -0.013 and Sig. = 0.086) and DDR and Number of Abstract's Word (r_s = -0.005 and Sig. = 0.527) were not significant.

Based on the findings of this study, several recommendations can be made. First, institutions should prioritize training for researchers and staff on the importance of comprehensive and accurate metadata creation. This could involve workshops or resources that emphasize best practices in metadata standards, particularly focusing on the Dublin Core schema. Second, regular audits of metadata completeness should be conducted to identify gaps and areas for improvement. By systematically reviewing and updating metadata records, repositories can enhance the discoverability of their academic outputs. Third, a comparison between different institutional repositories may be helpful for a better understanding of repository-specific biases, such as, age, size, and FAIRness.

The number of visits and downloads are two primary metrics used to evaluate the usage of dissertations (Alemneh & Phillips, 2011); however, they do not always align with more responsible metrics that accurately reflect user engagement. In this research, we proposed the DVR and DDR as more meaningful metrics for evaluation. These metrics provide a normalized measure of engagement, allowing for fair comparisons among dissertations across different departments and disciplines. By utilizing DVR and DDR, institutional repositories can gain deeper insights into how their resources are being accessed and utilized. We recommend that institutional repositories incorporate these two metrics into their reporting systems.

Many university libraries expend great energy to create Library of Congress Subject Headings (LCSH) for their university's ETDs. For future studies, it is recommended that researchers investigate the influence of subject headings, particularly LCSH, on the visibility of ETDs. While this study was unable to explore this aspect in depth due to the limited availability of LCSH—only 1,388 ETDs contained such headings—there remains a significant opportunity to assess how the presence and quantity of subject headings correlate with the number of views and downloads. Furthermore, it is noteworthy that among these 1,388 records, only 310 included abstracts, which have been identified as a significant predictor of engagement metrics.

Recognizing that submitters may not have the time to fill out all fields in the metadata schema, we recommend that institutional repositories embed tools to streamline the metadata creation process. Implementing metadata detection techniques that can automatically extract relevant information from the full text of dissertations could alleviate the burden on authors and ensure that essential metadata is captured. As institutions strive to foster open access and maximize research dissemination, embedding required tools and techniques in repositories, especially artificial intelligence-based tools, can help metadata practices. Recent studies (Oyighan et al., 2024) remind that there are still many challenges in the successful application of artificial intelligence could potentially aid in minimizing manual input, and Provenzano et al. (2024) that the scalability of artificial intelligence system allows for enhanced discoverability. However, it is clear that more research will be required to unpack use cases.

In conclusion, this study showed the key role of metadata completeness in enhancing user engagement within the DSpace@MIT repository. We found that well-structured and comprehensive

metadata significantly influences the visibility and accessibility of dissertations, ultimately impacting their retrieval and usage. This work serves as a call to action for the academic community to prioritize metadata quality, ensuring that research outputs are not only accessible but also effectively reach their intended audiences, thereby enriching the global knowledge ecosystem.

References

Adam, U. A., & Kaur, K. (2021). Institutional repositories in Africa: Regaining direction. *Information Development*, *38*(2), 166-178. https://doi.org/10.1177/02666669211015429

Alemneh, D., Donovan, B., Halbert, M., Han, Y., Henry, G., Hswe, P., McMillan, G., & (Lucy) Wang, X. (2014). *Guidance Documents for Lifecycle Management of ETDs* (M. Schultz, N. Krabbenhoeft, & K. Skinner, Eds.). Educopia Institute. https://educopia.org/wp-content/uploads/2018/07/Guidance_Documents_for_Lifecycle_Management_of_ETDs_0.pdf

Alemneh, D., & Phillips, M. (2016). Indexing quality and effectiveness: An exploratory analysis of electronic theses and dissertations representation. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1-4. https://doi.org/https://doi.org/10.1002/pra2.2016.14505301111

Alemneh, D. G. (2008). MAINTAINING QUALITY METADATA: TOWARD EFFECTIVE DIGITAL RESOURCE LIFECYCLE MANAGEMENT. In *Knowledge Management* (Vol. Volume 7, pp. 313-322). WORLD SCIENTIFIC. https://doi.org/doi:10.1142/9789812837578_0026

10.1142/9789812837578_0026

Alemneh, D. G., & Phillips, M. E. (2011, 26-30 June 2023). Assessing the Usage of Electronic Theses and Dissertations: An Overview of ETD Statistics and Metrics in the UNT Libraries. 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL),

Anil Hirwade, M. (2011). A study of metadata standards. *Library Hi Tech News*, *28*(7), 18-25. https://doi.org/10.1108/07419051111184052

Baudoin, P., & Branschofsky, M. (2003). Implementing an Institutional Repository. *Science & Technology Libraries*, *24*(1-2), 31-45. https://doi.org/10.1300/J122v24n01_04

Bruce, T. R., & Hillmann, D. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In *Metadata in Practice*. American Library Association.

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2024). *Evaluating the Linguistic Coverage of OpenAlex* [Working paper].

Chapman, J. W., David, R., & and Shreeves, S. A. (2009). Repository Metadata: Approaches and Challenges. *Cataloging & Classification Quarterly*, *47*(3-4), 309-325. https://doi.org/10.1080/01639370902735020

Chisale, A., & Phiri, L. (2023). *Towards Metadata Completeness in National ETD Portals for Improved Discoverability* 26th International Symposium on Electronic Theses and Dissertations (ETD2023), Gandhinagar, Gujarat, India. https://ir.inflibnet.ac.in/handle/1944/2412?mode=full

Choudhury, M. H. (2023, June 26–30, 2023). *ETDSuite: An Library for Mining Electronic Theses and Dissertations* JCDL'23, Santa Fe, New Mexico. https://www.cs.odu.edu/~cs_mchou001/website/resources/paper/JCDL_2023_DC-final.pdf

Choudhury, M. H., Salsabil, L., Jayanetti, H. R., Wu, J., Ingram, W. A., & Fox, E. A. (2023, 26-30 June 2023). MetaEnhance: Metadata Quality Improvement for Electronic Theses and Dissertations of University Libraries. 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL),

Comber, A. J., Fisher, P. F., Harvey, F., Gahegan, M., & Wadsworth, R. (2006). Using Metadata to Link Uncertainty and Data Quality Assessments. In A. Riedl, W. Kainz, & G. A. Elmes (Eds.), *Progress in Spatial Data Handling: 12th International Symposium on Spatial Data Handling* (pp. 279-292). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-35589-8_18

d'Aquin, M., Kirstein, F., Oliveira, D., Schimmler, S., & Urbanek, S. (2023). FAIREST: A Framework for Assessing Research Repositories. *Data Intelligence*, *5*(1), 202-241. https://doi.org/https://doi.org/10.1162/dint_a_00159

DCMI Usage Board. (2020). *DCMI Metadata Terms*. Dublin Core Metadata Initiative (DCMI). Retrieved Agust 20 from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5*(1), 31-49. https://doi.org/10.1162/qss_a_00286

Fear, K., & Donaldson, D. R. (2012). Provenance and credibility in scientific data repositories. *Archival Science*, *12*(3), 319-339. https://doi.org/10.1007/s10502-012-9172-7

Glogoff, S. J., & Forger, G. J. (2000). Metadata Protocols and Standards. *Internet Reference Services Quarterly*, *5*(4), 5-14. https://doi.org/10.1300/j136v05n04_03

Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics*, *88*(2), 653-661. https://doi.org/10.1007/s11192-011-0412-z

Kasonde, C. C., & Phiri, L. (2023). *Assessing and Promoting Metadata Quality for Electronic Theses and Dissertations in Institutional Repositories Using a Policy-Driven Approach* 26th International Symposium on Electronic Theses and Dissertations (ETD2023), Gandhinagar, Gujarat, India. https://ir.inflibnet.ac.in/handle/1944/2412?mode=full

Kenfield, A. S. (2019). Metadata Documentation Practices at ARL Institutional Repositories. *Libraries and the Academy*, *19*(4), 667-699. https://doi.org/10.1353/pla.2019.0041

Kumar, V., Chandrappa, & Harinarayana, N. S. (2024). Exploring dimensions of metadata quality assessment: A scoping review. *Journal of Librarianship and Information Science*, 09610006241239080. https://doi.org/10.1177/09610006241239080

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., & Manitsaris, A. (2008). *A Conceptual Framework for Metadata Quality Assessment*

MIT Libraries. (2024a). *About MIT Theses in DSpace@MIT*. MIT Libraries. Retrieved September 01 from https://libguides.mit.edu/dspace/about-theses

MIT Libraries. (2024b). *MIT Thesis FAQ: Thesis Checklist*. MIT Libraries. Retrieved September 01 from https://libguides.mit.edu/mit-thesis-faq/checklist

Musen, M. A., O'Connor, M. J., Schultes, E., Martínez-Romero, M., Hardi, J., & Graybeal, J. (2022). Modeling community standards for metadata as templates makes data FAIR. *Scientific Data*, *9*(1), 696. https://doi.org/10.1038/s41597-022-01815-3

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, *10*(2), 67-91. https://doi.org/10.1007/s00799-009-0054-4

Osman, R., K., Y. I. A. M., & A, A. (2023). Metadata matters: evaluating the quality of Electronic Theses and Dissertations (ETDs) descriptions in Malaysian institutional repositories. *Malaysian Journal of Library and Information Science*, *28*(1), 109-125. https://doi.org/10.22452/mjlis.vol28no1.7

Oyighan, D., Ukubeyinje, E. S., David -West, B. T., & Oladokun, B. D. (2024). The Role of AI in Transforming Metadata Management: Insights on Challenges, Opportunities, and Emerging Trends. *Asian Journal of Information Science and Technology*, *14*(2), 20-26. https://doi.org/10.70112/ajist-2024.14.2.4277

Paiva, C. E., Lima, J. P. d. S. N., & Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often [10.6061/clinics/2012(05)17]. *Clinics*, *67*(5), 509-513. https://doi.org/10.6061/clinics/2012(05)17

Park, E. G., & Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. *The Electronic Library*, *29*(3), 394-407. https://doi.org/10.1108/02640471111141124

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

Park, J.-R., Brenza, A., & Lu, C. (2015). A comparative analysis of metadata best practices and guidelines: issues and implications. *International Journal of Metadata, Semantics and Ontologies*, *10*(4), 240-260. https://doi.org/10.1504/IJMSO.2015.074751

Phiri, L. (2020). Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories. *International Journal of Metadata, Semantics and Ontologies*, *14*(3), 234-248. https://doi.org/10.1504/IJMSO.2020.112804

Pottier, P., Lagisz, M., Burke, S., Drobniak, S. M., Downing, P. A., Macartney, E. L., Martinig, A. R., Mizuno, A., Morrison, K., Pollo, P., Ricolfi, L., Tam, J., Williams, C., Yang, Y., & Nakagawa, S. (2023). Keywords to success: a practical guide to maximise the visibility and impact of academic papers. *bioRxiv*, 2023.2010.2002.559861. https://doi.org/10.1101/2023.10.02.559861

Provenzano, T., Fernandez, R., Deets, C., & Kirmis, D. (2024). *Using AI to facilitate discoverability and curation of the ASU Library repository collections: Report* (Using AI to facilitate discoverability and curation of the ASU Library repository collections, Issue.

Rasuli, B., Solaimani, S., & Alipour-Hafezi, M. (2019). Electronic Theses and Dissertations Programs: A Review of the Critical Success Factors. *College & Research Libraries*, *80*(1), 60-75. https://doi.org/10.5860/crl.80.1.60

Robson, B. J., & Mousquès, A. (2016). Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. *Environmental Modelling & Software*, *75*, 94-104. https://doi.org/https://doi.org/10.1016/j.envsoft.2015.10.007

Roosa, S. (2024, April 18, 2024). *Institutional Repository Usage Statistics (IRUS) at DSpace@MIT* Third Annual LyrOpen Fair, Virtual Conference.

Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics*, *110*(1), 243-251. https://doi.org/10.1007/s11192-016-2161-5

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720-1733. https://doi.org/https://doi.org/10.1002/asi.20652

Tani, A., Candela, L., & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, *49*(6), 1194-1205. https://doi.org/https://doi.org/10.1016/j.ipm.2013.05.003

Tarver, H., Zavalina, O., Phillips, M., Alemneh, D., & Shakeri, S. (2014). How Descriptive Metadata Changes in the UNT Libraries' Collections: A Case Study. *International Conference on Dublin Core and Metadata Applications*, *2014*. https://doi.org/10.23106/dcmi.952136407

Van Wyk, B., & Mostert, J. (2014). African Institutional Repositories as Contributors to Global Information: A South African Case Study. *Mousaion: South African Journal of Information Studies*, *32*(1), 98-114. https://doi.org/10.25159/0027-2639/1704

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R.,…Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18