Tips: All submissions must start by submitting an English extended abstract with a maximum of 800 words. The extended abstract should clearly deal with the problem/motivation/goal, the methodological approach, and anticipated results. The abstract MUST also include the title of the paper or poster, name(s) of the author(s), and full address information. For all formats, references do not count toward the page limit. Each submission will be reviewed in double-blind by at least two members of the program committee. All submissions must be original works, not previously published or under review for publication elsewhere, in English, in Microsoft Word format (DOC, DOCX), and in the APA 7th Edition Professional Paper format.

Presentation sub-themes:

- General Issues Related to ETDs.
- Global Visibility of ETDs.
- Institutional Repository Platforms.

Do More Complete Dissertations' Metadata Get More Engagement?

Behrooz Rasuli, Iranian Research Institute for Information Science and Technology (IranDoc), rasuli@irandoc.ac.ir; Michael Boock, Oregon State University, michael.boock@oregonstate.edu; Joachim Schöpfel, University of Lille, joachim.schopfel@uni-lille.fr; Brenda Van Wyk, The University of Pretoria, brenda.vanwyk@up.ac.za

Keywords

Electronic Theses and Dissertations, Metadata Standards, Open Science, Altmetrics, Bibliometrics.

Research Problem and Motivation

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.

Repositories employ various methods to ensure ETDs are described thoroughly with quality metadata. These range from requesting researchers to provide more comprehensive information during ETD deposit to policy-driven approaches (Kasonde & Phiri, 2023) and even proposals for automated quality improvement (Choudhury et al., 2023). While a universally agreed-upon definition of metadata quality remains elusive, 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.

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].

Research Objectives

The current research aims to:

- Determine the completeness of DSpace@MIT doctoral dissertations' metadata;
- Identify the number of views and downloads of ETDs within DSpace@MIT;
- Explore the relationship between ETD metadata completeness and the number of views within DSpace@MIT;
- Explore the relationship between ETD metadata completeness and the number of downloads within DSpace@MIT;
- Explore the relationship between the number of views and the number of downloads of ETDs within DSpace@MIT;
- Identify metadata fields within ETDs that exhibit consistently higher completeness rates compared to others.

Methodology

This study investigates the relationship between ETD metadata completeness and its impact on discoverability as measured by views and downloads within an IR through a bibliometrics method. DSpace@MIT, the institutional repository of the Massachusetts Institute of Technology, will serve as a case study. A random sample of 647 dissertations will be selected from DSpace@MIT. Stratified random sampling will be employed to ensure representation across different academic fields and years of publication within the collection. This approach balances the need for a representative sample while controlling for potential confounding factors like academic discipline or year of publication that might influence views and downloads. For each dissertation in the sample, the following data will be extracted: (1) ETD metadata: All available metadata fields associated with the dissertation will be collected; and (2) Usage statistics: The number of views and downloads for each dissertation, as reported by DSpace@MIT's Institutional Repository Usage Statistics (IRUS), will be recorded.

The completeness of ETD metadata will be evaluated by calculating the number of completed fields for each dissertation in the sample. Microsoft Excel will be used to automate this process by counting the number of populated fields within the extracted metadata. To account for potential variations in user interest across different academic disciplines and publication years, the number of views and downloads for each dissertation will be normalized. A suitable normalization technique will be selected based on the distribution of these variables. SPSS statistical software will be employed to explore the relationships between these variables: (1) ETD metadata completeness, (2) Number of views, and (3) Number of downloads. Appropriate

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

statistical tests, such as correlation analysis or regression analysis, will be chosen to examine the strength and direction of the relationships between these variables.

MIT was chosen as a case study for several reasons:

- High Visibility: As a renowned institute with international reach, MIT dissertations likely attract a significant number of views and downloads, providing a robust dataset for analysis.
- Controlled Variables: Focusing on a single institution allows for control over the "reputability of institute" variable, potentially mitigating its influence on the findings.
- 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 not only comprehensive metadata for dissertations but also readily accessible usage statistics for each document, facilitating data collection for both metadata completeness and discoverability measures.

By employing a controlled case study approach with robust statistical analysis, this research aims to contribute valuable insights into the relationship between ETD metadata completeness and discoverability within institutional repositories.

Anticipated Results

By analyzing the relationships between these factors, the study expects to shed light on the impact of metadata completeness on the discoverability of ETDs within an institutional repository setting. The expected outcomes are:

- The analysis will determine the average completeness level of ETD metadata within DSpace@MIT.
- The study will explore potential correlations between the completeness of ETD metadata and the number of views/downloads on the platform.
- A relationship between the number of views and downloads of ETDs might be identified.
- The research aims to identify metadata fields within ETDs that are consistently more complete than others.

References

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),

Kasonde, C., & Phiri, L. (2023, 11/2023). *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), Gujarat, India. http://ir.inflibnet.ac.in/handle/1944/2412

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

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

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