Assessing and Promoting Metadata Quality for Electronic Theses and Dissertations in Institutional Repositories Using a Policy-Driven Approach

Cecilia. C. Kasonde[1] and Lighton Phiri[2]

[1]Institute of Distance Education, The University of Zambia

[2] Department of Library and Information Science, The University of Zambia

Abstract

Electronic Theses and Dissertations (ETDs) have become integral to higher education institutions' (HEIs) repositories, facilitating accessibility, sharing, and preservation of research. High-quality metadata is crucial for effective ETD discovery, retrieval, and preservation. The Zambia National ETD portal project seeks to aggregate metadata from HEIs for a nationwide portal. This study examines ETD metadata quality and ingestion processes in HEI repositories, focusing on institutional repository policies. Using mixed methods, the research assesses ETD metadata in Zambian HEIs with functional repositories. Quantitative analysis rates metadata completeness using the ETD-MS standard. Interviews with 16 stakeholders from 8 HEIs identify policy-related influences on metadata quality. Result shows low ETD metadata completeness, averaging 6.43 out of 10. Institutions often diverge from standards like ETD-MS, creating their own or adapting existing ones. Compliance with NDLTD's 14 metadata elements is poor, with optional elements like contributors and coverage often missing. The study highlights metadata's importance, particularly completeness, for IRs. It advocates prioritizing metadata quality through standardized practices, creator training, and quality control mechanisms. Addressing metadata enhances ETD discoverability and utility both nationally and globally.

Keywords: Electronic Theses and Dissertations (ETD), Institutional Repository Policies, Metadata Completeness, Metadata Quality

1. Introduction

Electronic theses and dissertations (ETDs) have revolutionized the way academic research is disseminated and accessed in higher education institutions (HEIs). With the emergence of institutional repositories (IRs), ETDs have found a secure and accessible home. These repositories play a vital role in preserving scholarly output, facilitating ease of sharing, and increasing the visibility of research contributions. However, the success of IRs in achieving these goals hinges on the quality of metadata associated with the ETDs they host.

Corresponding Author: Cecilia C. Kasonde, Email: kasondececiliac@gmail.com.

Metadata is the crucial information about a digital resource that enables effective discovery, retrieval, and management. In the context of ETDs, metadata includes essential details such as title, author, abstract, keywords, subject classifications, and more. High-quality metadata enhances discoverability, aids browsing, and ensures long-term preservation, contributing significantly to the overall effectiveness and usefulness of IRs.

The current research highlights the results and implications of a study that delves into the challenges of metadata quality for ETDs in institutional repositories. The research aimed to assess the existing metadata quality in all HEIs in Zambia, explore factors affecting metadata quality, and identify policy-centric approaches to promote metadata excellence.

2. Related Work

2.1 Metadata

Metadata is data about data and is essential for digitized objects and electronic archiving, particularly in higher learning institutions where the focus has been on archiving theses and dissertations. The Dublin Core metadata standard, developed in 1995 in Dublin, Ohio, popularized the term, but the concept of cataloguing in libraries goes back in history (Wright, 2007).

Metadata plays a vital role in digital libraries and facilitates search, evaluation, acquisition, and use of resources (IEEE, 2001). It includes basic elements like title, author, and year of publication, as well as more comprehensive information such as technical features, copyright properties, and annotations.

Metadata is crucial for understanding information in data warehouses and XML-based web applications (Smith, 2007). It ensures the survival and accessibility of resources in the future, aids in accurate searching and retrieval, and helps evaluate resources. Additionally, metadata assists in managing, maintaining, and preserving digital collections, supports interoperability, and ensures the security and authentication of digital resources.

Ultimately, metadata acts as the link between information creators and users, enabling efficient access to scholarly research and educational resources, considering diverse cultural and lingual contexts (IEEE, 2001). Using metadata in accordance with international standards is key to achieving these objectives. (Fox et al., 2001).

2.2 Problems with Metadata Quality

The majority of today's metadata is produced by untrained individuals working alone without sufficient support, leading to a wide range of quality issues and lack of interoperability (Hillmann, et al, 2004). Many digital libraries handle big data sets, prioritizing efficiency, and automation over user interaction (Suleman, 2012). Quality errors in metadata and full-text object data can hinder access to online documents (Beall, 2006). The quality of metadata in digital collections and repositories is often found to be incomplete,

ambiguous, and inconsistent (Stvilia et al., 2004). Challenges in metadata creation are evident in learning object repositories and open ePrint archives (Barton, Currier & Hey, 2003).

For instance, DSpace was criticized for lacking essential descriptive metadata components for Electronic Theses and Dissertations (ETDs) (Park, 2009). In the Dryad Repository, data quality problems were identified with metadata elements, impacting text mining and data analysis (Rousidis et al., 2014). The process of creating high-quality metadata often relies on human annotators, which becomes limiting as the number of digital resources increases (Palavitsinis, 2013).

Overall, problematic metadata quality is a significant challenge faced by various cross-domain repositories. Automated methods can only provide partial solutions, and quality assurance processes are essential to improve metadata accuracy and usability.

2.3 Mechanisms That Facilitate Ingestion of High-Quality ETDs.

Academic libraries use various processes and procedures for creating and distributing metadata for electronic theses and dissertations (ETDs). A survey in 2017 analyzed ETD metadata policies, workflows, and practices among 137 public and private institutions in the United States. The survey identified patterns and differences in metadata creation, policy-setting, and access, which were found to be unique to each institution to suit their specific needs.

Libraries have integrated ETDs into their institutional repositories, enhancing access beyond traditional catalogues. This inclusion benefits students and universities by promoting graduate education, expanding research, increasing visibility, and educating stakeholders about digital technology. However, managing metadata for ETDs across different platforms presents challenges in cataloguing procedures and automation.

Metadata quality is crucial for effective use of materials in institutional repositories. Empirical research by Chassanoff (2009) showed that institutions often have manual metadata quality checking processes, with limited use of automated tools. User feedback also plays a role in discovering quality issues. Institutions maintain documentation on metadata policies, but challenges persist as they adapt to managing digital materials.

Metadata is expected to perform diverse functions beyond access and retrieval, including situating resources in social and historical contexts. Research by Phiri (2020) proposed a method for automatically categorizing ETDs using machine learning techniques to improve metadata quality in institutional repositories.

A nationwide survey (Park & Tosaka, 2010) revealed that MARC, AACR2, and LCSH are commonly used metadata schemas and controlled vocabularies. Dublin Core (DC) and EAD are also widely used. Collection-specific considerations and existing technological infrastructure influence the selection of metadata schema and controlled vocabularies.

However, metadata interoperability remains a challenge due to locally created metadata and the lack of shareable mechanisms for extensions and variants.

Metadata is integral to the lifecycle of ETDs, institutional repositories, and digital libraries. Addressing challenges and embracing new technologies can enhance the effectiveness and accessibility of ETD metadata for researchers and users.

3. Methodology

To assess the metadata quality, data was collected from the institutional repositories. The metadata records were examined for the presence or absence of specific elements recommended by the National Digital Library of Theses and Dissertations (NDLTD) for comprehensive description of Electronic Theses and Dissertations (ETDs). The completeness score was calculated by dividing the number of present elements by the total recommended elements and then multiplying by ten.

3.1 Sample

All 62 registered HEIs under the Higher Education Authority (HEA) in Zambia. 53 Private HEIs and 9 Public HEIs.

3.2 Instrument

Data was collected using an online questionnaire and online interviews. An online questionnaire was used in order to determine HEIs that have IRs. Once the HEIs had been identified and their IR URLs verified, the researcher harvested data from the IRs for analysis. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) was used to harvest metadata The analysis involved getting metadata that had been prepared and measuring its completeness.

3.3 Design

A mixed-method approach was employed. A quantitative analysis of ETD metadata from all HEIs with functional IRs was conducted, by focusing on metadata completeness—the completeness metric was arrived at by ascribing scores to individual metadata elements, relative to the ETD-MS metadata standard. In order to identify factors that affect metadata quality, interviews were conducted by 16 key stakeholders, from 8 HEIs, involved in drafting IR ingestion policies and, additionally, individuals involved in the ingestion of ETDs into IRs.

4. Results

4.1 State of metadata quality in HEI IRs

According to Figure 1, the results reveal the completeness scores for the metadata elements in the IRs. The scores range from 3.57 to 6.43 on a scale of 10, indicating the level of metadata quality compliance across the institutions.

Figure 1: Metadata Elements in IRs

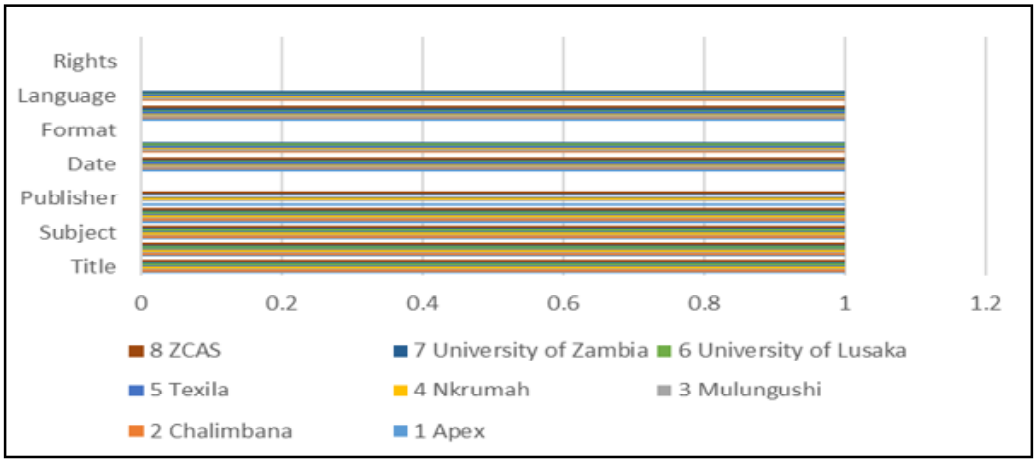

Table 1 indicates the completeness scores which were calculated as follows:

Completeness score = (Number of present elements / Total recommended elements) * 10

Sum of completeness scores: 5.00 + 5.71 + 6.43 + 6.43 + 6.43 + 5.71 + 6.43 + 5.00 = 47.14

Total number of institutions: 8

Average completeness score = Sum of completeness scores / Total number of institutions Average completeness score = 47.14 / 8 = 5.89 (rounded to two decimal places)

Therefore, the average completeness score for all the institutions is approximately 5.89 out of 10.

Table 1 : Completeness Score for HEI IRs

Institution	Completeness Score
Apex	5.00
Chalimbana	5.71
Mulungushi	6.43
Nkrumah	6.43
Texila	6.43
University of Lusaka	5.71
University of Zambia	6.43
ZCAS	5.00

Based on the completeness scores calculated for the metadata elements in the HEI IRs, we can see that the scores range from 3.57 to 6.43 out of a maximum score of 10. This indicates that the overall metadata quality in HEI IRs varies from moderate to relatively high.

However, there is still room for improvement as none of the institutions achieved a completeness score of 10.

Analysing the completeness scores, the highest score of 6.43 was achieved by the University of Zambia, indicating a relatively higher level of metadata quality compared to the other institutions. The presence of a higher number of metadata elements in the University of Zambia's IR suggests a more comprehensive and accurate description of their ETDs.

For the other institutions, the completeness scores ranged from 3.57 to 5.00. This shows that while there is a moderate level of metadata quality, improvements can be made in certain areas. The missing metadata elements, such as Contributor, Format, Coverage, Rights, and Thesis. Degree, contribute to the lower scores and highlight areas that require attention.

The results indicate that the completeness of ETD metadata is extremely low, with the highest average completeness score being 6.43, out of a total score of 10. Further analysis of the results indicates a lack of adherence to international standards such as ETD-MS and, additionally, a focus on specific metadata elements such as dc_contributor/Advisor", "dc_coverage" and "dc_thesis. degree". Interview sessions conducted with stakeholders revealed that institutions dealing with ETDs have all developed their own standards or adapted existing metadata standards.

All institutions attempt to describe the author, work and content in which the work was produced in a way useful to both researchers and library staff maintaining the work in its electronic form. The international organization NDLTD provides a standard set of 14 metadata elements for ETDs. However, the study revealed low compliance with the ETD-MS standard, with many optional metadata elements missing, such as contributor, format, coverage, rights, and thesis degree. These elements are important for describing ETDs, including the contribution of supervisors/advisors.

Addressing these missing elements in the IRs of all the institutions would significantly enhance the metadata quality. Including the Contributor element would acknowledge the intellectual contributions of individuals or organizations, providing proper attribution. The Format element would ensure better understanding of the technical characteristics of the ETDs, enabling compatibility and usability. The Coverage element would provide contextual information about the scope of the research, enhancing relevance. The Rights element would ensure compliance with legal and ethical considerations, promoting responsible usage. Lastly, the Thesis. Degree element would add credibility and context to the ETDs, enabling assessment of academic achievements.

Therefore, while the overall metadata quality in HEI IRs shows promise, there is still a need to address the missing metadata elements and strive for higher completeness scores. By doing so, HEIs can improve the discoverability, accessibility, and usability of their ETDs, ensuring accurate and comprehensive representation of scholarly resources in their institutional repositories.

It is worth noting that the missing elements vary across institutions, suggesting inconsistencies in metadata practices and standards. Efforts should be made to address the missing elements and enhance the completeness and consistency of metadata in HEI IRs.

4.2 Challenges of Metadata Quality

The study revealed a striking discrepancy in the quality of ETD metadata across HEIs in Zambia. The completeness of metadata elements was found to be disappointingly low, with the highest average completeness score being only 6.43 out of a total score of 10. This inadequacy directly impacts discoverability and retrieval of ETDs, hindering researchers from accessing valuable academic content.

Moreover, the lack of adherence to internationally recognized metadata standards, such as ETD-MS, poses additional challenges. Many institutions had either developed their own metadata standards or adapted existing ones, leading to inconsistencies and omissions in crucial metadata elements. The study also found low compliance with important metadata fields, such as contributor information, format, coverage, rights, and thesis degree, which are vital for comprehensive descriptions of ETDs.

4.3 Institutional Repository Policies and Metadata Quality

To address the issues of metadata quality, the study explored the role of institutional repository policies. A policy-driven approach could play a significant role in promoting metadata excellence and uniformity across all HEI IRs. However, the research revealed that currently, the emphasis on metadata quality in repository policies is limited.

Incorporating standardized metadata practices, providing training and support for metadata creators, and implementing quality control mechanisms are essential steps that institutions should take to improve metadata quality. Adhering to established metadata standards, like ETD-MS, and encouraging compliance with all essential metadata elements should be embedded in repository policies to enhance the overall discoverability and accessibility of ETDs.

5. Discussion

Metadata quality dimensions encompass completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility (Bruce & Hillmann, 2004). The study specifically focuses on completeness.

The results of the study assessing the completeness of ETD metadata in institutional repositories are concerning and shed light on the challenges faced in maintaining high-quality metadata for electronic theses and dissertations as also observed by Tani, A etal (2013), they discussed the challenges that digital libraries face in maintaining high-quality metadata, such as inconsistencies, incompleteness, and inaccuracies that can hinder resource discovery and retrieval.

Phiri, L. (2020) recommends the use of automated or machine learning-based classification methods to enhance the metadata quality of ETDs. He discussed how these techniques can assist in assigning appropriate metadata descriptors, keywords, and subject classifications to ETDs, leading to improved organization and discoverability.

The highest average completeness score of 6.43 out of 10 indicates a significant gap in providing comprehensive metadata for ETDs, limiting their discoverability and accessibility. One of the critical issues highlighted in the study is the lack of adherence to international standards, particularly the ETD-MS standard. Chassanoff, A (2009) supports the researchers' findings in his study by suggesting that metadata activities may not yet be streamlined into an institution's workflow and organizational structure. The lack of formalized quality control and procedures seem to indicate that metadata quality is an after-thought for most institutions.

The ETD-MS standard, established by the Networked Digital Library of Theses and Dissertations (NDLTD), provides a set of 14 metadata elements that are essential for describing ETDs. However, the research found low compliance with this standard, with many optional metadata elements missing. These missing elements include crucial information such as contributor details (supervisors/advisors), format, coverage, rights, and thesis degree, which are vital for providing a comprehensive and accurate representation of the ETDs.

The study's findings also reveal a tendency among institutions to develop their own metadata standards or adapt existing ones. Park, J & Tosaka.Y, (2010) The leading criteria in selecting metadata and controlled-vocabulary schemata are collection-specific considerations, such as the types of resources, nature of the collection, and needs of primary users and communities. Existing technological infrastructure and staff expertise also are significant factors contributing to the current use of metadata schemata and controlled vocabularies for subject access across distributed digital repositories and collections.

While this approach may be driven by the desire to cater to specific institutional needs, it results in a lack of uniformity and consistency in metadata practices across repositories. This inconsistency hampers the interoperability of ETD metadata and poses challenges for aggregating and integrating ETDs into national and global ETD services.

The study revealed that six out of eight institutions lack IR policies, indicating a deficiency in structured guidelines for managing their IRs. Furthermore, the absence of an ingestion policy across all institutions underscores the absence of formalized guidelines for incorporating digital content such as ETDs. The absence of both suggests potential inconsistencies in managing digital objects among institutions, thereby influencing metadata practices, file formats and quality assurance. This deficiency can impact accessibility, preservation and the overall effectiveness of the repositories. However, the study's interview results revealed the need for institutional policies.

Palmer, Lauren, and Newton (2008) also support a mandatory policy for IR that contains a clear-cut explanation of Intellectual Property Rights, copyrights, and other legal concerns and more common factors need to be covered.

To address these issues, it is essential for institutions to prioritize the adoption of standardized metadata practices, such as the ETD-MS standard, in their repository policies. Emphasizing the inclusion of all essential metadata elements and ensuring compliance with international standards will significantly enhance the overall quality of metadata in institutional repositories. This, in turn, will improve the discoverability and accessibility of ETDs, benefiting researchers, librarians, and other stakeholders in the academic community.

Furthermore, the study's interviews with stakeholders involved in dealing with ETDs highlight the importance of describing authors, works, and content in a manner that is useful to both researchers and library staff. Metadata serves as a bridge between ETDs and their potential users, making it crucial to provide accurate and detailed information. A concerted effort to train and support metadata creators in adhering to standardized practices will lead to more consistent and informative metadata descriptions.

6. Way Forward

In light of the study's findings, it is imperative that HEIs in Zambia and beyond prioritize the enhancement of metadata quality in their IRs. A concerted effort to adopt best practices, international standards, and clear policies will help ensure that ETDs are optimally described, enabling efficient discovery and access to valuable research output.

In the rapidly evolving digital landscape, where research dissemination and access are increasingly reliant on robust metadata, addressing the challenges highlighted in this study becomes a strategic priority. Collaboration among institutions, leveraging the expertise of librarians, and engaging researchers and other stakeholders in the process will foster an environment where high-quality metadata becomes the norm rather than the exception.

7. Conclusion

Effective management of ETDs in institutional repositories requires a concerted effort to address metadata quality comprehensively. The study's results underscore the need for HEIs to recognize the critical role of metadata in the discoverability and accessibility of academic research. By implementing policy-driven approaches, supporting standardized metadata practices, and investing in the training of stakeholders, institutions can elevate the quality of metadata associated with ETDs, enhancing their scholarly impact and facilitating knowledge dissemination worldwide.

The study highlights the need for institutions to prioritise metadata quality in their repository policies and workflows. Implementing standardised metadata practices, providing training and support for metadata creators, and establishing quality control mechanisms are essential steps towards promoting metadata quality in HEI IRs.

The research underscores the critical role of high-quality metadata in facilitating effective searching, browsing, and long-term preservation of electronic theses and dissertations. By promoting the adoption of international metadata standards, encouraging compliance with all essential metadata elements, and fostering collaboration

among institutions, the academic community can work towards enhancing the discoverability and accessibility of ETDs. This, in turn, will contribute to the broader dissemination and impact of scholarly research in the digital age

References

Barton, J., Currier, S., & Hey, J. (2003). Building quality assurance into metadata creation: An analysis based on the learning objects and e-prints communities of practice. In DCMI (Ed.), Proceedings of the International Conference on Dublin Core and Metadata Applications (pp. 39-48). https://dcpapers.dublincore.org/pubs/article/view/770

Beall, J. (2006). Metadata and Data Quality Problems in the Digital Library. University of Colorado at Denver and Health Sciences.

Chassanoff, M. A. (2009). Metadata Quality Evaluation in Institutional Repositories: A Survey of Current Practices. A Master's Paper for the M.S. in I.S degree. University of North Carolina at Chapel Hill. https://cdr.lib.unc.edu/indexablecontent/uuid:4e5644a0-4135-4edd-93b1-337e2c99a73f

Fox, E., & Marchionini, G. (2001). Special issue on digital libraries. Communications of the ACM, 44(5).

Hillmann, D. I., Dushay, N., & Phipps, J. (2004). Improving Metadata Quality: Augmentation and Recombination. In DCMI (Ed.), Proceedings of the International Conference on Dublin Core and Metadata Applications. https://dcpapers.dublincore.org/pubs/article/view/770

IEEE. (2001). Draft standard for learning object metadata, draft 6.1, April 2001. http://ltsc.ieee.org/wg12/index.html

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in The Digital Age. portal: Libraries and the Academy, 3(2), 327–336. https://doi.org/10.1353/pla.2003.0039

Palavitsinis, N. (2014). Metadata Quality Issues in Learning Repositories (PhD thesis). University of Alcala de Henares. http://www.slideshare.net/nikospala/metadata-quality-issues-in-learning-repositories

Palmer, C. L., Teffeau, L. C., & Newton, M. P. (2008). Strategies for institutional repository development: A case study of three evolving initiatives. Library Trends, 57(2), 142-167.

Park, J., & Tosaka, Y. (2010). Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability. https://www.semanticscholar.org/paper/Metadata-Creation-Practices-in-Digital-Repositories-Park-Tosaka/406b174d7da10c7bec0ff31913e891bbb8b85ef2

Park, J., & Lu, C. (2009). Metadata professionals: Roles and competencies as reflected in job announcements, 2003–2006. Cataloging & Classification Quarterly, 47, 145–160.

Phiri, L. (2018). Research visibility in the global South: towards increased online visibility of scholarly research output in Zambia. In Proceedings of the IEEE International Conference in Information and Communication Technologies. http://dspace.unza.zm/handle/123456789/572

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

Smith, I. (2007). Metadata, metadata schemes and metadata standards. IGBIS seminar 'Digital library standards and metadata the basics.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S. L., & Cole, T. W. (2004). Metadata quality for federated collections. In Proceedings of ICIQ04—9th International Conference on Information Quality (pp. 111-125). Cambridge, MA.

Suleman, H. (2012). The NDLTD Union Catalogue: Issues at a global scale. In Proceedings of the 15th International Symposium on Electronic Theses and Dissertations. Retrieved from https://repositorioacademico.upc.edu.pe/handle/10757/622568

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/10.1016/j.ipm.2013.05.003