Towards Metadata Completeness in National ETD Portals for Improved Discoverability

Adrian Chisale[1] **and Lighton Phiri**[1]

[1]Department of Library and Information Science, University of Zambia, Lusaka, Zambia

Abstract

Zambia has 61 registered Higher Education Institutions (HEIs) with the Higher Education Authority. In order to facilitate easy access to Electronic Theses and Dissertations (ETDs), works are underway to implement a National ETD portal. The diversity of the ETD source IRs poses a challenge in ensuring that ETD metadata is of high quality, with respect to the completeness, adversely affecting discoverability of ETD digital objects. This paper outlines a study conducted to identify HEIs with functional IRs; empirically assess the relative quality of ETD metadata from HEI IRs and investigate potential ways of identifying sources of missing metadata elements. Questionnaires were distributed to 61 HEIs in order to identify HEIs with functional and interoperable IRs. In addition, ETD metadata from HEIs with functional IRs was harvested using the Open Archives Initiative for Metadata Harvesting protocol and subsequently analyzed in order to assess the metadata completeness. Finally, a combination of document analysis of policy documents and, additionally, content analysis of randomly sampled ETD manuscripts from HEIs with functional IRs was conducted in order to identify potential sources of missing metadata. Out of 61 HEIs, only 10 (16.1%) of HEIs in Zambia had implemented functional IRs. The analysis of ETD metadata indicates that there is generally non-compliance of ETD metadata to the ETD-MS metadata standard, the de facto metadata schema for ETDs developed by the Networked Digital Library of Theses and Dissertations. In order to identify reliable sources of missing metadata elements, content analysis of policy documents was performed, alongside an analysis of randomly sampled ETD bitstreams. Potential sources of missing metadata from the ETD bitstream make it possible for automated extraction techniques to be employed to automatically generate missing metadata elements.

Keywords: Electronic Theses and Dissertations (ETD), Institutional Repositories, Metadata Quality

1. Introduction

High-quality digital object metadata has been in used and demonstrated to facilitate core bibliographic functions of discoverability, use, provenance, currency, authentication, and administration (Park, 2009). While the need for adequate metadata is necessary for the various types of digital objects, the increased rate at which Electronic Theses and Dissertations (ETDs) are being generated requires that much emphasis is placed on ETD metadata as well. This need is further necessitated by the increase in the number of

Corresponding Author: Lighton Phiri, Email: lighton.phiri@unza.zm

downstream aggregation services that harvest ETD metadata at scale. These downstream aggregation services include services that harvest ETD metadata at a global scale, such as the Networked Digital Library of Theses and Dissertations (NDLTD)'s Union Catalog[1] and the Open Access Theses and Dissertations portal[2], and, additionally, services that are localized to a region or country. Downstream aggregation services localized to countries, such as South Africa's National ETD Portal[3], enable the aggregation of ETD metadata from various Higher Education Institution (HEI) Institutional Repositories (IRs) via a single unified interface (Webley et al., 2011).

Due to the diverse nature of the sources of ETD metadata, one of the major challenges experienced by downstream services is the poor quality of metadata, relative to completeness and correctness. Suleman reports that "errors in incoming records are a recurring problem" during the harvesting of metadata by the NDLTD Union Catalog (Suleman, 2012). As a way of addressing metadata quality issues, (Tani et al., 2013) described four main approaches for addressing metadata quality issues: implementation of metadata guidelines, standards, and application profiles; evaluation strategies for identifying metadata issues; semi-automated metadata generation; and metadata pre-processing and augmentation approaches.

The implementation of the Zambian National ETD Portal project, which is a nationwide project, aims at aggregating ETD metadata from HEIs in the Republic of Zambia (Phiri, 2018). The large quantities of metadata, originating from heterogeneous sources necessitate the need for effective ways of searching and browsing for content in the National ETD portal by ensuring that the ETD metadata harvested from external sources is of high quality. Hence, the need to conduct a survey to identify important commonly missed metadata elements during ingestion of ETDs in Higher Education Institutions (HEIs) Institutional Repositories (IRs) and, additionally, an investigation into how missing metadata elements can potentially be automatically generated.

Paper outlines work done to identify important commonly missed metadata elements during ingestion of ETDs in Higher Education Institutions (HEIs) Institutional Repositories (IRs) and, additionally, an investigation into how missing metadata elements can potentially be automatically generated.

2. Objectives

The main objective of this study is to comprehensively evaluate the quality of ETD metadata, relative to completeness, associated with ETDs originating from HEIs in Zambia and, additionally, to explore how missing metadata elements could potentially be automatically generated.

The specific objectives are as follows:

- To identify Higher Education Institutions that has functional and interoperable online Institutional Repositories in Zambia.
- To determine the relative quality of ETD metadata originating from HEIs in Zambia.
- To investigate the sources of missing metadata elements for the automated generated using NLP techniques

3. Related works

An Institutional Repository (IR) is a collection of digitalized items that contain metadata that helps to safeguard and make readily accessible the information developed by a particular educational institution. IRs acts as a medium through which knowledge is shared to the general public, at the same time provides opportunity for both scholars and institution to visible at the world market which in return helps to maintain good reputations online (Clobridge, 2010). The general status of institutional repositories (IRs) in African nations remains in stagnation. As a result, the performance of established institutional repositories in African nations generally remains below par, despite the potential of worldwide open access to research in few nations in the global south, a situation that can be linked to low adoption of IRs from onset (Phiri, 2018; Yusuf, Ifijeh, & Emmanuel, 2019; Adam 2021).

The "Implementation of Zambian National Electronic Theses and Dissertations (ZANETD) portal, which aims at aggregating ETDs from all Tertiary Institutions (HEIs) in Zambia," can only be successful if metadata ingesters provide high quality metadata. Hence, a need to survey in order to find out the current status of Institutional Repositories in all Universities and colleges in Zambia with a view of finding solutions on how the missing Electronic Theses and Dissertations Metadata could be generated for the national ETD portal (Phiri, 2018).

4. Metadata Quality in Institutional Repositories

One of the primary responsibilities of institutions dedicated to collecting and preserving information resources is the development of accurate and reliable metadata. The maintenance, preservation, presentation, and dissemination of digital objects are as crucial as the production of high-quality metadata. As a result, this essential task requires proper planning and resources. In a study, Baca (2008) indicated that "access to information does not increase with digitization." This is with a realization that digitalization alone without high-quality generated descriptors is not adequate to enhance resources accessibility, comprehensibility and usage to data and information consumers.

The increase in the development of new metadata standard with intent to manage information has sparked discussions of quality amongst information professionals. The endeavor has been the adoption of metadata as a remedy to information overload as it was recognized to be a gateway that helps consumers finds what they are looking for even in times of indecision (Bruce & Hillmann, 2004; Alemneh, 2009). As a result, metadata ingesters need to be concerned with the mechanics of producing high-quality metadata at all levels of their operations because the usefulness and worthiness of institutional repositories are all linked to the completeness of the metadata. Therefore, it can be deduced that, the poor the quality of metadata, the compromised the search capability of the repository. However, when the metadata is based on sound resource analysis, it increases the value of a resource (Ochoa, 2013).

The interoperability of Institutional repositories, at regional, national and at global level, can be hindered by the poor quality of ETD metadata. Weagley, Gelches, and Park (2010) in a study pointed out that "metadata must be comprehensive enough to incorporate all possible queries from the user and it should be discoverable in an aggregated environment.” It is metadata that determines the operation and interoperability of Institutional Repositories. To be effective, the bibliographic records between repositories must contain comparable fields and valid values for successful interoperability. This also entails that data entry processes must be standardized and that content descriptions follow accepted standards.

The need to create exceptional metadata has intensified. As a result, metadata influences how digital items are found and used. Benchmarks have been set by information specialists on classification and rating of metadata. Park (2009) observed that “metadata’s usefulness is measured by its capacity to carry out the essential bibliographic tasks of finding, usage, provenance, currency, authentication, and administration.” Therefore, metadata should be thorough and detailed such that the user may fully comprehend the function and content of the listed resource without actually visiting it. This is with the understanding that distributed search could negatively be affected when one of the connected repositories has substantially incomplete metadata. Hence, high-quality metadata is essential as the usefulness of a digital repository is strongly dependent upon comprehensiveness of the metadata that describes its resources.

Ochoa (2009) came up with some of the quality measurement metrics, which comprised “completeness, correctness, and consistency and full access capacity to individual local objects and connectivity to the parent local collection(s) were also used to determine the completeness of a metadata record.” This entails that completeness metrics does not only refer to assigning a large number of elements with values that describe an object but also through distributed search. Accuracy centers on the surrogacy of a particular item’s descriptive analysis like typos, layout, and subject matter involving the item under. In this study therefore, the measurement of the completeness of metadata was done by verifying the capability of having full access to individual local objects and connection to the parent local collection of ETD metadata and assigning of enough elements to the digital objects.

4.1 Metadata Elements of Digital Objects

“Descriptive, administrative, and structural metadata are various types of metadata in the repository which are used to manage information resources” (Riley, 2017). As it has been alluded from the above that metadata are crucial in preserving and displaying digital items, it is prudent to have all types of metadata in all digital objects in good quality. This is only possible when the right metadata standards are adopted with respect to the type of digitals objects to be managed in the IRs.

“There is no universal standard for all information resources, every metadata schema chosen determines the bits of metadata to be included and excluded” (Baca, 2008). However, there are two types of metadata schemata, which include specialized and generalized. Specialized systems demand more work and understanding to implement, yet the digital objects are described very well. While general scheme is characterized with minimal effort and knowledge with insufficient description to items. As a result, there has been wide range of metadata schemas created to accommodate diversity of information resources and disciplines. In case of Electronic Theses and Dissertations, the management and exchange of ETD metadata at a worldwide level was made possible by the creation of the ETD-MS metadata standard, which eventually became the de-facto metadata standard used to define ETDs in Institutional repositories. Therefore, the quality measurements in this study will also be based on the compliance of HEIs to ETD-MS.

5. Methods

Questionnaires were distributed to 61 HEIs in Zambia to identify HEIs that offer postgraduate programs and, additionally, to determine HEIs with functional and interoperable IRs. Metadata from HEIs with functional IRs were harvested using the Open Archives Initiative Protocol for metadata Harvester (OAI-PMH) and subsequently analyzed in order to assess the metadata completeness during ingestion, relative to the ETD-MS metadata standard (Lagoze et al., 2002; Hickey, Pavani & Suleman, 2010)

Documents analysis and semi structured interviews were also used to collect data. During content analysis, policy documents were analyzed in relation to prescribed standard outline and format in which ETD should be presented. While the ETDs were analyzed to ascertain the location of metadata on the preliminary pages.

The saturation principle was used to determine the sample for analysis in this study. This means that the total number of documents for analysis could not be known before a study begins. Therefore, sample size was established based on data saturation. The threshold of data redundancy was necessary to determine a representative sample. “This stage is reached when no new data is gathered and researchers are no-longer gaining new insights” (Merriam & Tisdell, 2016). Due to the uniform nature of Theses, which have a set structure and layout that is accepted by all higher education institutions, the sample size of electronic theses and dissertation metadata records (ETDs) for inclusion in content analysis was determined using the saturation principle. Thus, the saturation wasn’t proven until no new information emerged which came after an analysis of 25 metadata entries from all HEI repositories (Bowen, 2009).

Microsoft Excel was used to analyze data on HEIs that offer postgraduate programs to determine HEIs with functional and interoperable IRs. A combination of document analysis of policy documents and, additionally, content analysis of randomly sampled ETD manuscripts from HEIs with functional IRs was conducted to identify potential sources of missing metadata elements. The document analysis from policy documents was done to identify the standard guidelines on how Theses and Dissertations must be produced with regard to outline, structure, and information to be included on preliminary pages. On the other hand, content analysis on ETDs was conducted to identify the metadata elements that each repository used in the description of ETDs in every repository.

6. Results

6.1 Higher Education Institutions that have functional and interoperable online Institutional Repositories in Zambia

Zambia has 61 Higher Education Institutions registered with the Higher Educational Authority (Higher Education Authority, 2022). Out of 61 Universities 43 responses were received. Further investigations from the other 18 registered institutions indicated that they were not operational, and some were closed. Therefore, the analyses of responses only included complete responses. Consequently, all incomplete responses were deleted. Multiple responses from the same institutions were equally removed and only remained with unique ones per university. Therefore, this analysis was based on unique and complete responses only.

Results indicated that out of the 43 HEIs in Zambia, 28 offered postgraduate programmes, with a mere 10 HEIs that installed IRs. In addition, all IRs from HEIs used DSpace as the Institutional repository platform.

A survey in 43 HEIs to ascertain the higher educational institutions with functional and interoperable online institutional repositories found that 10 (23.3%) of HEIs in Zambia had implemented IRs and 33 (76.6%) of HEIs were still managing their ETDs in analogue format. Out of 10 that implemented and adopted IRs in Zambia 3 were found to be non-functional and 1 (2%) did not have ETDs content despite offering postgraduate programs at the time of data collection and analysis of results as shown in the figure 1 below.

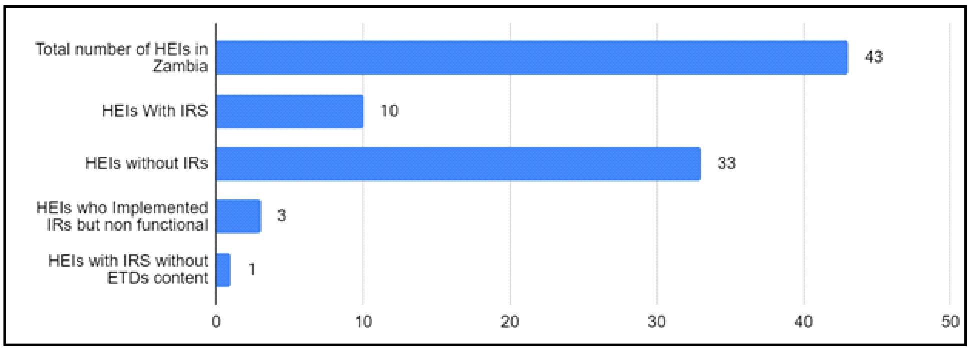

Figure 1: Status of institutional Repositories in Zambia

Among the functional institutional repositories, it was found that eight (8) IRs representing 80% do not share metadata and only two (2) representing 20 % activated the Open archive initiative protocol (OAI) for sharing metadata as shown in Figure 2 below.

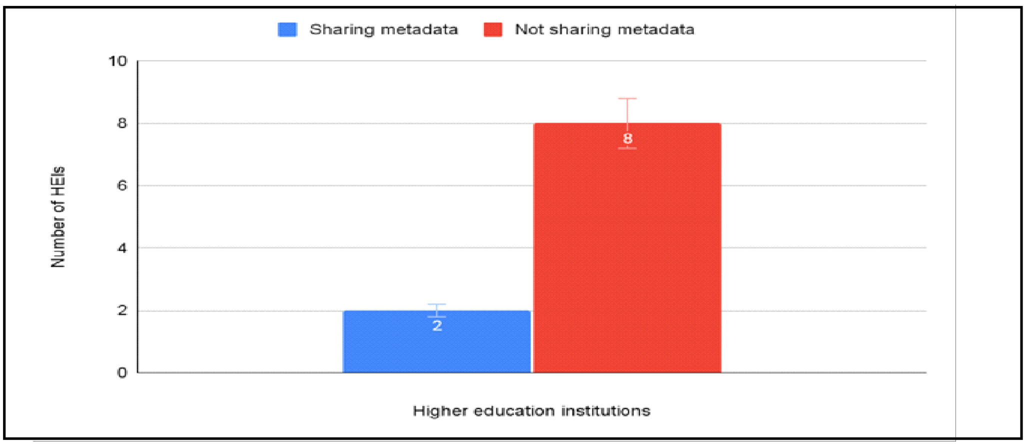

Figure 2: Interoperability of Institutional Repositories in Zambia

6.2 Relative quality of ETDs metadata in Higher Education Institutions in Zambia

The research discovered that there were ten (10) HEIs that adopted and implemented IRs in Zambia. Among these HEIs, only eight (8) had functional Institutional Repositories that preserves digital contents in various subject areas. Digital repositories were all selected for metadata harvesting, and out of selected 8 repositories, only University of Zambia and University of Lusaka were found active for metadata harvesting. Between the two Universities that activated the OAI PMH, research found that only UNZA IR had its records indexed, making it to be the only repository which was compliant with Open Archive Initiative protocol for metadata harvester (OAI PMH). However, metadata quality analysis in Higher Learning Institutions was not limited to HEI repositories that activated Open Archive Initiative Protocol for metadata harvesting (OAI PMH).

6.3 Situation analysis of metadata quality in Institutional Repositories in Zambia

Over 10,000 records were harvested from the University of Zambia Institutional Repository. After conducting data cleaning, it was found that, at a time of data collection i.e., (23rd February 2022), UNZA IR contained only 4003 records of thesis and dissertations. Out of the 4003 records, it was found that 757 records had their title elements missing, 845 records had creator elements missing, 814 records had subject elements missing and 910 records had description elements missing as shown in figure 3 below.

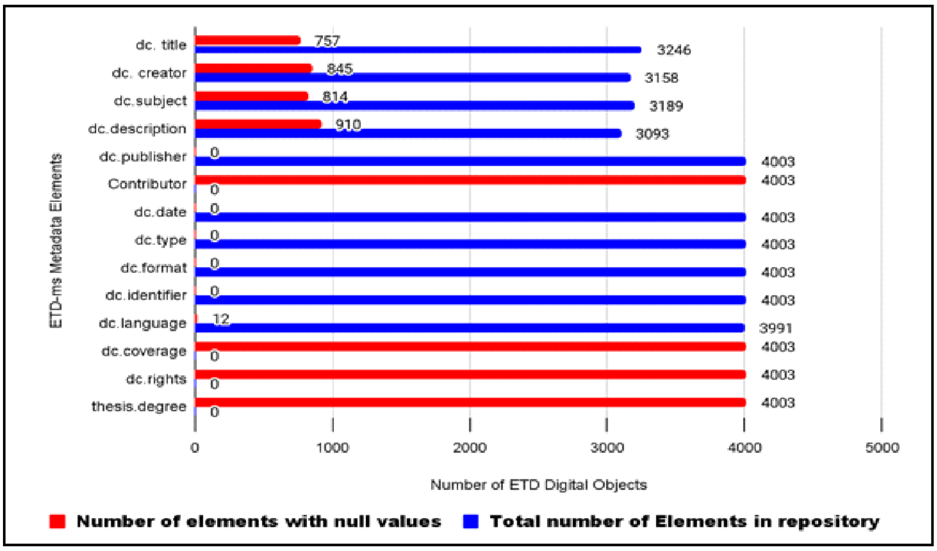

Figure 3: Metadata elements used at UNZA IR

The above figure shows that the University of Zambia institutional repository uses nine (9) Dublin core elements to describe Electronic Theses and Dissertations (ETDs) against de facto ETD-ms standard schema's requirements. These elements include: dc.title, dc.creator, dc.subject, dc.description, dc.publisher, dc.date, dc.type, dc.identifier and dc.language.

6.4 University of Lusaka Institutional Repository

The Electronic Theses and Dissertations metadata from the University of Lusaka (UNILUS) were harvested and analysed to identify the metadata elements used to describe ETDs. The results reviewed that, UNILUS Institutional Repository uses seven (7) elements that comprised dc.creator, dc.date, dc.identifier, dc.Language, dc.subject, dc.title, and dc.type as shown in figure 4 below.

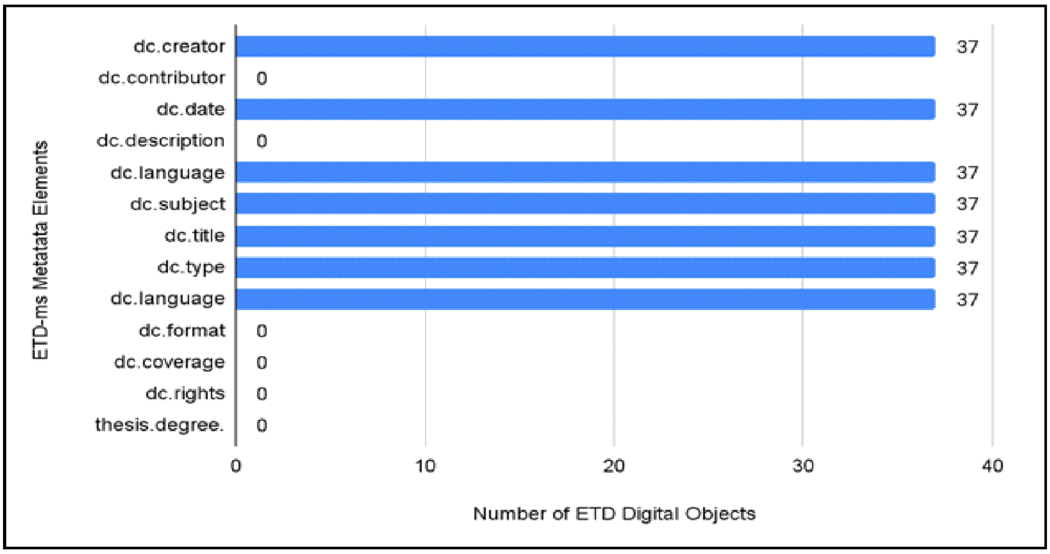

Figure 4: Metadata elements used at UNILUS Vs ETD-MS

UNILUS ETDs Metadata were further analysed to determine the relative quality using completeness metrics and their compliance to ETD-MS, the de facto metadata for ETDs. The study discovered that UNILUS had 37 ETD records. All the 37 bibliographic records in the repository had five (5) ETD-MS metadata elements missing. The missing elements included; dc.contributor, dc.format, thesis.degree, dc.coverage and dc.rights.

A similar analysis was equally done on Cavendish University, Mulushownhi university, Zambia Institute of Accountancy studies University, Chalimbana University Institutional Repository and Lusaka Medical Apex University Institutional Repositories, the results showed that all Institutional repositories were not ETD-MS compliant as show in the table 1 below.

Table 1: Metadata Elements used in Higher Education Institutions

metadata elements used in HEIs in Zambia	Higher Education Institutions				
	Cavendish University	Mulungushi university	Zambia Institute of Accountancy studies University	Chalimbana University Institutional Repository	Lusaka Medical Apex University
dc.creator	√	√	√	√	√
dc.contributor	x	x	x	x	x
dc.date	√	√	√	√	√
dc.format	x	√	x	x	x
dc.identifier	√	√	√	√	√
dc.description.	√	√	√	√	√
dc.coverage	x	√	x	x	x
dc.language	√	√	√	√	x
dc.publisher	√	√	√	x	√
dc.subject	√	√	√	√	√
dc.title	√	√	√	√	√
dc.type	√	√	√	√	√
dc.rights	x	x	√	x	x
thesis.degree	x	x	x	x	x

The policy documents and ETDs manuscripts analyzed indicated potential sources of missing metadata elements from IRs. For instance, while supervisor/advisor details were found on the title page of most manuscripts for most HEIs, these details were not available in manuscripts for some HEIs as shown in the table 2 below:

Table 2: Sources of Missing Metadata in Institutional Repositories

HEIs	Metadata element missing	Location of missing metadata in ETD (Sections of ETD)
University of Zambia IR	Supervisor / Contributor/ advisor	Acknowledgements/ certificate of approval
Mulungushi University IR	Supervisor / Contributor/ advisor	Title page/ Supervisor's Recommendation/ Acknowledgement
Chalimbana University IR	Supervisor / Contributor/ advisor	Approval/ Acknowledgement/ Declaration
Cavendish University IR	Supervisor / Contributor/ advisor	Title page/ Declaration/ Acknowledgement/

Lusaka Apex Medical University IR	Supervisor / Contributor/ advisor	Title page/Supervisor's declaration
University of Lusaka IR	Supervisor / Contributor/ advisor	Title page/ Certificate of Approval/ Acknowledgement
Zambia Centre for Accountancy Studies University I R	Supervisor / Contributor/ advisor	Title page/ Acknowledgement

The identification of sources of information to be used to automatically generate missing metadata presents opportunities for the implementation of effective distributed services in the country. As part of current and future work, Natural Language Processing models are being developed to automatically generate missing metadata elements.

7. Conclusion

The research revealed that there were eleven (11) Higher education Institutions (HEIs) with functional Institutional Repositories that preserve digital contents in various subjects in Zambia. Digital repositories were all selected for metadata harvesting, and out of selected 11 repositories, only University of Zambia and University of Lusaka were found active for metadata harvesting. Between the two Universities that activated their Open Archive Initiative Protocol for Metadata Harvester (OAI PMH), only UNZA IR had its records indexed, making it to be the only Repository, which was compliant with Open Archive Initiative Protocol for metadata harvester (OAI PMH).

The metadata quality analysis of all ETDs in Zambia were found to be compromised. This was observed in the missing metadata elements, elements with missing values and non-compliant to ETD-MS, the de facto metadata schema for electronic theses and dissertations. As a result, the current state of metadata from HEIRs, is said to greatly affect the downstream services both local and at global level.

The feasibility of automatically generating ETDs missing metadata in Higher Education Institutional Repositories is possible through the natural language processing Libraries like SpaCy Library using Named-entity recognition (NER) which can be used to extract missing elements from all higher education institutional repositories in Zambia.

References

Adam, U. (2021). Institutional repositories in Africa: Regaining direction. Available on https://journals.sagepub.com/doi/abs/10.1177/02666669211015429

Alemneh, D. (2009). Proceedings of the American Society for Information Science and Technology 46(1), http://digital.library.unt.edu/ark:/67531/metadc29318/

Baca, M. (2008). Introduction to metadata. 3rd ed. Getty Research Institute: Los Angeles

Bowen, G.A. (2009). Document Analysis as a Qualitative Research Method. Qualitative Research Journal,9 (2), 27-40. https://doi.org/10.3316/QRJ0902027

Bruce,T.R. & Hillmann, D. I. (2004). The Continuum of metadata quality: defining, expressing, exploiting. Available at https://hdl.handle.net/1813/7895.

Clobridge, A. (2010). Building a digital repository program with limited resources. Sawston. Cambridge. doi:10.1533/9781780630458.

Hickey, T., Pavani, A., & Suleman, H. (2010). ETD-MS v1.1: An interoperability metadata standard for electronic theses and dissertations. Networked Digital Library of Theses and Dissertations. https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html#

Higher Education Authority. (2022). Higher Education Authority. Higher Education https://doi.org/10.1177/02666669211015429

Lagoze, C. et al., (2002). Open archives initiative protocol for metadata harvesting. Available on http://www.openarchives.org/OAI/openarchivesprotocol.html

Merriam, S. B., & Tisdell, E. J. (2016). Qualitative research: a guide to design and implementation (4th ed.). San Francisco, CA: Jossey Bass.

Ochoa, X., (2013). Automatic evaluation of metadata quality in digital repositories. International Journal of digital libraries, (10), 67–91. https://doi.org/10.1007/s00799-009-0054-4.

Park, J. R. (2009). Metadata quality in digital repositories. a survey of the current state of the art. cataloging & classification quarterly. https://doi.org/10.1080/01639370902737240

Phiri, L. (2018). Research visibility in the global South: towards increased online visibility of scholarly research output in Zambia. In IEEE International Conference in Information and Communication Technologies. http://dspace.unza.zm/handle/123456789/5723

Riley, J. (2017). Understanding metadata, what is metadata, and what is it for. National Information Standards Organization. https://www.fidgeo.de/fileadmin/user_upload/

Suleman, H. (2012). The NDLTD Union Catalog: Issues at a Global Scale: Standard for Electronic Theses and Dissertations. Available on https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/622568/ETD2012

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

Weagley, J., Gelches, E., & Park, J.-R. (2010). Interoperability and metadata quality in digital video repositories: A study of Dublin Core. Journal of Library Metadata, 10(1), 37–57. https://doi.org/10.1080/19386380903546984.

Webley, L., Chipeperekwa, T., & Suleman, H. (2011). Creating a National Electronic Thesis and Dissertation Portal in South Africa. 14th International Symposium on Electronic Theses and Dissertations. International Symposium on Electronic Theses and Dissertations, Cape Town. http://pubs.cs.uct.ac.za/id/eprint/748/1/etd2011_webley.pdf

Yusuf, F. & Ifijeh, G. Emmanuel, O. (2019). Institutional Repositories in Africa: Issues and Challenges. 10.4018/978-1-5225-8437-7.ch008.