Identification of Sources for Missing Electronic Theses and Dissertations Metadata in Higher Education in Zambia

By

Adrian Chisale

A dissertation submitted to the University of Zambia in partial fulfilment of the requirements of the degree of Master of Library and Information Science

THE UNIVERSITY OF ZAMBIA

LUSAKA

2024

Declaration

I, Adrian Chisale, declare that this dissertation is a product of my effort and has not been presented to any academic institution certification or degree.

Sign: …………………………………………

Date: …………………………………………

Certificate of Approval

This dissertation by Adrian Chisale has fulfilled the requirements for the award of the Master's degree in Library and Information Science by the University of Zambia.

Examiner's Name   Signature   Date

……………………….. . ……………………   … …………………….

…………………….………   …………………….   …......………………

……………………………   ……………………   ………………………

Abstract

Low-quality metadata can lead to vagueness, poor recall, and inconsistent search results, underscoring the importance of robust quality assurance mechanisms in Information Retrieval systems. This study addresses the challenge of ensuring the completeness and quality of Electronic Theses and Dissertations (ETDs) metadata within Higher Education Institutions (HEIs) in Zambia. Despite the presence of 61 registered HEIs in Zambia, efforts to establish a National Electronic Theses and Dissertations (ETDs) Portal are underway to enhance accessibility to these scholarly documents. However, the diversity of ETD sources poses a challenge in maintaining metadata integrity, thereby impacting the discoverability of ETD digital objects.

This research aims to identify HEIs with functional Institutional Repositories (IRs), assess the relative quality of ETD metadata from these repositories, and explore methods to address missing metadata elements. A questionnaire was distributed to all 61 HEIs to ascertain the presence of functional and interoperable IRs. Additionally, ETD metadata from HEIs with functional IRs were harvested using the OAI-PMH protocol and analyzed to evaluate metadata completeness. Furthermore, a combination of document analysis of policy documents and content analysis of randomly sampled ETD manuscripts from HEIs with functional IRs was conducted to identify potential sources of missing metadata. The findings indicate that out of 61 HEIs, only 10 (16.39%) had implemented functional IRs. Moreover, the analysis of ETD metadata reveals non-compliance with the ETD-MS metadata standard, established by the Networked Digital Library of Theses and Dissertations. Variations in the location of certain ETD-MS metadata elements across HEIs were observed. Identified sources of missing ETD-MS metadata elements from IRs present opportunities for automated extraction techniques to generate missing metadata, thereby facilitating the successful implementation of the Zambia National ETD portal and improving the visibility of ETDs in Zambia.

Keywords: *Electronic Theses and Dissertations (ETDs), Institutional Repositories (IRs), Metadata completeness, ETD-MS metadata standard, Metadata quality measurements, Dublin Core metadata schema.*

Dedication

This research is dedicated to my mother, Doreen Juliya Moose and my late father Mr Benson Muundwe Chisale who engineered my academic journey without conditions attached. Lastly to my wife, Mercy Nyanga and my children, Adriel, Ellah, Ephraim A and Peggy for their love and care amidst challenges. To you all I say let us keep on pushing for the betterment of our family.

Acknowledgements

Table of contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

Abbreviation	Description
AGETDM	Automatic Generation of Electronic Theses and Dissertations Metadata
CRISP	Cross Industry Standard
HEIs	Higher Education Institutions
ETDs	Electronic Theses and Dissertations
IRs	Institutional Repositories
MS	Metadata Standard
NLP	Natural Language Process
OAI-PMH	Open Archive Initiative Protocol for Metadata Harvesting
DCMI	Dublin Core Metadata Initiative
UNZA	University of Zambia
CHAU	Chalimbana University
ZCASU	ZCAS University
CUZ	Cavendish University
MU	Mulungushi University
LAMU	Lusaka Apex Medical University
CBU	Copperbelt University
DRGS	Directorate of research and graduate studies

CHAPTER ONE: INTRODUCTION

1.1   Overview

This section outlines the context of the study, articulates the problem statement, defines the study's purpose, states research objectives, poses research questions, and underscores the significance of the study.

1.2   Introduction

Higher Education Institutions (HEIs) in Zambia have a vital mandate of providing training with certifications ranging from Certificate, Diploma, Degree, Master, and PhD degree.[1] As part of the requirement, every postgraduate student produces a research paper in form of dissertation or theses before they are conferred with degrees. These academic publications must be available to the intended user community (Higher Education Authority, 2019). Theses and dissertations constitute a significant portion of university information resources, serving as the foundational elements for research collections and essential components for sustainable research endeavours. Higher Education Institutions have recently established IRs to share their research contributions with the global community. Electronic Theses and Dissertations (ETDs) have emerged as crucial and advantageous in developing countries like Zambia. The adoption of ETDs in Zambia began in the last decade of the 21st century. Since then, their popularity has steadily increased among HEIs and the research community. This trend is driven by the desire to disseminate newly generated knowledge promptly (Wani, 2015)

The transition from traditional paper formats to the electronic submission of theses and dissertations by numerous HEIs has led to the establishment of ETD systems. Many HEIs either develop their own ETD systems or acquire commercial ETD software to manage scholarly work, i.e., storage, preservation, and dissemination. The growing complexity and emergence of new technologies in contemporary times

---

[1]   https://hea.org.zm/wp-content/uploads/2021/06/Higher-Education-Amendment-Act-No.-23-of-2021.pdf

have transformed the expectations of library users, their demands now extend beyond traditional boundaries, encompassing remote access to full-text resources, including theses and dissertations. In order to meet these evolving user needs, there is a need for a continuous enhancement in metadata quality. The advent of downstream services, which harvest metadata for ETDs to create global digital libraries, poses a significant challenge to metadata quality. The realisation of the dream of global digital libraries hinges on addressing this challenge, ensuring that ETDs from all participating institutions can be easily accessed.

1.3   Background to the Study

The quality of metadata plays a crucial role in enhancing the discoverability of a record and facilitating effective search and retrieval processes. Ochoa and Duval (2009) observed that many implementations of IRs adopted a lenient stance toward quality assurance despite the widespread metadata integration in resource cataloguing within digital repositories. This leniency often stems from the assumption that metadata is crafted by subject matter experts or professional cataloguers, implying an inherent level of acceptable quality. However, experts in a specific field may need to gain expertise in metadata creation. Additionally, hiring professional indexers for resource cataloguing proves impractical for most repositories due to scalability constraints and associated costs. Electronic Theses and Dissertations (ETDs) subpar metadata quality can detrimentally impact downstream services. Hence, there is a need to put proper measures to safeguard the metadata generation process and, in some cases, even develop and deploy intelligent systems that would automatically generate missing metadata in response to challenges experienced by downstream services (Suleman, 2012).

The integration of technology, open-source initiatives, and the evolving paradigm of intellectual property rights has fostered the development of ETDs projects. A notable outcome of this trend is the establishment of the Networked Digital Library of Theses and Dissertations (NDLTD), an international organisation dedicated to advancing the creation, accessibility, and preservation of ETDs from academic institutions globally.

The NDLTD, as of the year 2021, offered access to an extensive collection, surpassing a million ETDs (NDLTD Union Archive, 2022). The NDLTD is an international organisation with a dedicated mission to encourage the adoption, creation, utilization, dissemination, and preservation of ETDs. It actively supports electronic publishing and advocates for open access to scholarly content to facilitate global knowledge sharing (NDLTD, 2022). Similarly, the NDLTD was founded in 1995 to foster a collaborative approach in this domain. Its primary objectives include developing benchmark software programs, tools, systems, standards, and guidelines. The organization serves as a platform for discussions on various issues, particularly metadata harvesting (Suleman and Fox, 2012). Presently, NDLTD comprises over 140 member institutions, mainly universities, which have collectively devised and implemented ETD-MS, an interoperable metadata standard designed explicitly for ETDs.

The focus of NDLTD extends beyond the mere retrieval, browsing, access, downloading, and reading of publications. It also encompasses their creation, submission, and all associated workflow activities. For digital libraries to be both cost-effective and scalable, it is essential to implement appropriate automation and ensure that the use of the digital library aligns with standard community activities. Fox and McMillan (2004) noted that NDLTD has played a pivotal role in advocating for the Open Access movement. This initiative, rooted in automation and open sharing of metadata, which gave rise to the Open Archives Initiative (OAI). The OAI facilitated the development and utilization of repositories, each supporting a repository access protocol known as the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). Operating as a client-server protocol based on the Hypertext Transfer Protocol (HTTP), OAI-PMH enables incremental metadata transfer among networked systems. Its simplicity and generality make it suitable for application in diverse contexts and communities. Suleman & Fox (2012) pointed out that OAI-PMH allows the Union Archive to collect metadata about ETDs from Universities, Colleges and National libraries that are ETD-Ms compliant through aggregators worldwide.

1.4 Statement of the Problem

In Zambia's higher education landscape, the establishment of Institutional Repositories (IRs) designed to house Electronic Theses and Dissertations (ETDs) faces multifaceted challenges, particularly regarding their functionality, interoperability, and the quality of associated metadata. While certain academic institutions have embarked on initiatives to develop such repositories, the extent to which these platforms effectively operate and seamlessly interact with one another across the academic landscape remains uncertain [2](Chan, 2008). This lack of clarity undermines the repositories' potential to facilitate scholarly collaboration and broaden access to research outputs. Moreover, the varying quality of metadata accompanying ETDs within these repositories poses significant hurdles in terms of information retrieval and the dissemination of scholarly knowledge (Suleman, 2012). Factors contributing to these challenges include incomplete submissions by librarians, technical limitations within repository systems, inconsistent adherence to metadata standards, complications arising from data migration processes, and occasional errors introduced during the metadata creation process (National ETD Portal, 2018; Phiri, 2018). Addressing these issues necessitates a comprehensive understanding of the sources of missing ETD MS metadata in ETD manuscripts, coupled with strategic interventions aimed at enhancing repository functionality, improving interoperability between institutional platforms, and standardising metadata practices across the academic sector. By doing so, stakeholders could foster a more robust scholarly ecosystem conducive to effective knowledge dissemination and academic collaboration throughout Zambia's higher education landscape.

1.5 Research Objectives

The main objective of this research was threefold,

The general objective was to   identify potential sources of missing ETD metadata IRs in Zambia. In this regard the study specifically sought to:

i. Identify Higher Education Institutions that have functional and interoperable IRs in Zambia.

---

[2] https://informationr.net/ir/13-1/paper333.html

ii. Determine the relative quality of ETD metadata in HEIs in Zambia.

iii. Identify potential sources of missing ETD metadata originating from HEIs in Zambia.

1.5.1 Research Questions

i. Which Higher Education Institution implemented Institutional Repositories that are functional and interoperable in Zambia?

ii. What is the relative quality of Electronic Theses and Dissertations Metadata in Higher Education Institutions in Zambia?

iii. What are sources of missing Electronic Theses and Dissertations metadata in Higher Education Institutions in Zambia?

1.6 Limitations of the Study

This study focused exclusively on Higher Education Institutions (HEIs) that offer postgraduate programs, specifically targeting those with functional IRs. Lack of comprehensive and updated information from HEA on all registered HEIs and their contact details, negatively affected the inclusion of some HEIs in the survey.

1.7 Significance of   the Study

The findings of this study hold significance for achieving interoperability across IRs and contributing to the development of a standardised application profile for Zambian IRs, this is in tandem with Franzl, Wanzenböck and Berger (2023) who observed that interoperability is a key requirement for the digitalisation projects IRs. The study's results could be vital in putting quality assurance mechanisms during metadata ingestion in IRs in Zambia. The results of this research could contribute to a successful establishment of a National Research Portal that consolidates all ETD metadata in Zambia. This assumption in agreement with Suleman (2012) who observed that generation of quality of ETD metadata is one of the factors that makes downstream services to succeed. It is also hoped that results of this study would improve the metadata quality that may enhance the visibility of local research resources world over.

1.8 Conceptual Framework

1.8.1 Cross-Industry Standard Process for Data Mining Model

This research was guided by the Cross-Industry Standard process for the data mining model (CRISP-DM). CRISP-DM is an open standard process framework model for data mining project planning (Huber, Wiemer, Schneider & Ihlenfeldt, 2019). It has six phases that includes: Business understanding, data understanding, Data preparation, modelling, evaluation and deployment. In this study only however, three phases were utilised in this study due to study setting.

1.8.1.1 Business Understanding

This phase aims to define the objectives and understand the preliminary requirements of the project. The phase highlighted the scope and skills required to complete the project by assessing risks and contingencies and conducting a cost-benefit analysis while selecting technologies and tools with well-defined detailed plans for each project phase.

1.8.1.2 Data Understanding

1.8.1.2.1 Data Sources

ETD metadata records from all functional HEI repositories in Zambia and postgraduate policy documents were data sources for analysis and interpretation. In contrast, primary data was obtained from Librarians from HEIs through interviews and questionnaires.

1.8.1.2.2 Data Collection

Metadata records were harvested using OAI-PMH and manually downloading ETD metadata from repositories that did not configure OAI-PMH protocol. Data on standard outline and location of missing metadata was collected using interviews and desk research. Desk research was done by analysing policy documents containing postgraduate regulations.

1.8.1.3 Data Preparation

This was a critical stage in research as it was determinant of the correct/ wrong interpretation of findings. The data preparation activities involved identifying and rectifying errors, such as removing duplicate records, handling missing values, standardising formats, filtering or faceting record information, and renaming the record columns. The cleaned data was then exported to a Google spreadsheet for another data processing.

1.9 Operational Definitions of Concepts

The following terms in the study will mean:

i. **Descriptive Metadata:** Descriptive metadata constitutes information that characterises the asset, including the resource title, authorship, and relevant keywords. It encompasses any data describing the resource that can be utilised for subsequent identification and discovery. Descriptive metadata is particularly robust as it offers multiple avenues for describing digital resources (Chapman, Reynold and Shreeves, 2009).

ii. **Dissertation**: A dissertation is an extensive scholarly writing document research conducted throughout a doctoral program. Awarding a Ph.D. typically follows the submission and defense of the dissertation (Enago Academy, 2021).

iii. **Electronic Theses and Dissertations (ETDs)**: An Electronic Thesis or Dissertation (ETD) is a document that explicates the research of a graduate student and expresses in a form simultaneously suitable for machine archives and worldwide retrieval. (Chapman, Reynold, &. Shreeves, 2009).

iv. **Higher Education Institutions**: Refers to public or private college or University which provides learning programs leading to award of certificates, diploma, bachelor's degrees and post graduate qualifications and conducts research and development in various academic disciplines (Zambia Higher Education Act no. 23 of 2023)

v. **Institutional Repositories (IRs):** These refers online archives designed to collect, preserve, and disseminate digital copies of the intellectual output of an institution, particularly a research institution (Yusuf, Ifijeh, Emmanuel, 2019.

vi. **Metadata Quality:** This encompasses seven general characteristics: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. It reflects the overall reliability and effectiveness of metadata (Park, 2009)**.**

vii. **Metadata**: Metadata is structured information that describes, explains, locates, or otherwise facilitates an information resource's retrieval, use, or management—often referred to as "data about data" or "information about information" (National Information Standards Organization, 2010).

viii. **Open Archives Initiative Protocol for Metadata Harvesting:** is a low-barrier mechanism facilitating repository interoperability. Data Providers expose structured metadata via OAI-PMH, and Service Providers make OAI-PMH service requests to harvest that metadata (Suleman, 2012).

ix. **Structural Metadata:** This defines relationships between digital objects and facilitates the effective presentation of digital assets when rendered. It indicates how a digital asset is organi**s**ed, such as the arrangement of pages in a book or the structure of notes in platforms like Evernote or OneNote. Structural metadata also helps navigate and present information in electronic resources (Riley, 2017).

x. **The Networked Digital Library of Theses and Dissertations:** NDLTD is an international organization that promotes the adoption, creation, use, dissemination, and preservation of electronic theses and dissertations (Suleman 2012).

xi. **Thesis**: a supervised and assessed scholarly work produced by the students pursuing post graduate studies preferably a PhD level (Park & Richard, 2011).

1.10 Summary of Chapter One

Chapter one focused on the background information of the study on institutional repositories and metadata generations in higher education institutions in Zambia which is critical in managing scholarly works. The chapter has shown that comprehensiveness of metadata plays a vital role in effective search and visibility of ETDs online as they have broad implications on downstream services that

automatically harvest ETD metadata. The Chapter has shown the need for IRs to be compliant to ETD-MS, the de facto standard established by NDLTD in management ETDs. The next chapter provided a review of the literature relevant to the study of institutional repositories and metadata generations in higher education institutions in Zambia.

CHAPTER TWO: LITERATURE REVIEW

2.1 Overview

This section delved into the literature pertinent to the current study. A literature review proves invaluable as it allows a researcher to discern how peers have addressed comparable issues. It unveils sources that might otherwise remain unknown, offering an opportunity to evaluate one's work from historical and comparative standpoints. Moreover, a literature review aids researchers in gaining insights into study conduct, identifying potential pitfalls to avoid, and discerning effective new methodologies. It serves as a critical foundation for contextualising and refining the research approach.[3] Therefore, the scope of this chapter was guided by the following research objectives:

i.   Identify Higher Education Institutions that have functional and interoperable IRs in Zambia.

ii.   Determine the relative quality of ETD metadata in HEIs in Zambia.

iii.   Identify potential sources of missing ETD metadata originating from HEIs in Zambia.

2.2 Institutional Repositories

The IRs are digital archive belonging to an institution. It is a tool for collecting, storing and disseminating information to advance scholarly communication. The HEIs and other research entities globally reconsiders the production and usage of information resources in the digital format and they have been looking for ideal means for capturing and reusing the intellectual output for teaching and research (Betz & Hall, 2015). Globally, IRs are catalytic of transformative role in academic knowledge management (KM). This role is vital to knowledge management activities including the acquisition, creation, conversion, sharing, dissemination, transfer, preservation, and reuse of knowledge in higher education (Arjun, 2017). The content of IRs for HEIs includes research journal articles (preprints/post prints) undergoing peer review, and digital versions of theses and dissertations. It might also include

---

[3]https://pubrica.com/academy/research/why-is-it-important-to-do-a-literature-review-in-research/

other digital assets such as administrative documents, course material, etc. IRs could purely contain scholarly, administrative, teaching and research materials, both published and unpublished.

2.2.1 Institutional Repository platforms

Institutional Repositories were first developed as an online solution for development, administration and disseminating of the scholarship from all HEIs. The repositories quickly transformed into platforms for publishing and showcasing the holdings of a given Institution. The Added support for multimedia, brought greater depth to repository collections. Since 2000, several repository platforms have been developed, each with their own set of benefits and technical criteria. Therefore, every organisation needs an IR platform capable of hosting, managing, and showcasing their research output. Amongst the types, the broader class includes open-source platforms and proprietary platforms. Since every HEIs has unique needs, demands, and requirements, the choice for the platform is dependent upon the platform that works best for them (UNESCO, 2014).

2.2.1.1 Proprietary Software

This is a commercial platform that the user needs to purchase for use. Any maintenance, upgrade and editing are solely dependent on the commercial developer which always comes at huge cost due to lack of options. Examples of commercial IRs software includes:

i. **Digital Commons by bepress (Elsevier):** Digital Commons is a popular commercial platform for institutional repositories. It offers a range of features for managing and showcasing scholarly works, including articles, theses, dissertations, and other digital collections. It provides customisation options, analytics, and integration with various library systems.[4]

ii. **Symplectic Elements (Clarivate Analytics):** Symplectic Elements is a research information management system designed to help institutions

---

[4] https://bepress.com/products/digital-commons/why-digital-commons/

manage and display their research outputs. It integrates with various data sources to automatically populate researcher profiles, publication lists, and institutional repositories. It offers customization options and support services.[5]

iii. **ProQuest Institutional Repository (ProQuest):** ProQuest offers a commercial solution for institutional repositories, providing tools for hosting, managing, and disseminating scholarly content. It includes features such as customizable workflows, metadata management, and analytics to track usage and impact.[6]

iv. **Elsevier Pure (Elsevier):** Elsevier Pure is a research information management system that helps institutions manage and display their research outputs. It offers features for collecting, organizing, and showcasing scholarly works, including integration with institutional repositories for seamless dissemination of research outputs.[7]

v. **CONTENTdm (OCLC):** CONTENTdm is a digital collection management software that enables institutions to create and manage digital repositories for various types of materials. It provides tools for metadata management, digital preservation, and online access to institutional collections.[8]

These proprietary solutions offer a range of features and services to meet the needs of institutions for managing and disseminating scholarly works and digital collections. Institutions should consider factors such as features, customisation options, pricing, and support services when evaluating these options (Velmurugan, 2010).

The main weakness of this software include, lack of large community of support as the owner could be an individual developer or an organization. This is a setback due to lack of continuity in times of uncertainties with the developers. The unavailability

---

[5] https://www.symplectic.co.uk/theelementsplatform/
[6] https://about.proquest.com/en/dissertations/
[7] https://www.elsevier.com/products/digital-commons
[8] https://www.oclc.org/en/contentdm.html

of source code entails that the users are unable to modify the Software on their own other than depending upon the developer and this comes with huge costs. Despite not open for modifications, these platforms have dedicated developers who respond to faults and makes sure that the software meet user specifications with optimum security (UNESCO, 2014).

2.2.1.2 Free and Open-Source Software

The free and open-source software (FOSS) implies non-commercial platforms that are both free software and open source. their licenses grant users the rights to use, share, study, modify and update its version due to availability of its source code. This approach has gained both momentum and acceptance as the potential benefits have been increasingly recognised by both individuals and corporations as free software. Some of examples of these platforms includes; DSpace, EPrints, Invenio, Islandora, Museolog, Omeka, Refbase, RefDB, SobekCM and Greenstone (UNESCO, 2014).

Velmurugan (2010) pointed out that FOSS are cost effective due to their flexibility to customisation and elimination of licensing that reduces the overall project costs. This is ideal for HEIs with limited budgets. The availability of source code allows to modify the software to meet the specific needs for an individual organisation. The community collaborations are some benefits that comes with FOSS. The FOSS belong to a community of vibrant developers who are experts that helps to improve the software both in services and security.

Open-Source software however, have their own weaknesses among them is the lack of dedicated support despite belonging to the community of experts unlike the commercial platforms that runs on business concepts. This software also has some hidden costs making then not 100% free. All organisations that implement FOSS suffers from costs related to support, implementation, and security. At the same time some FOSS may have limitations on usage when applied to a particular entity. However, the benefits of using OSS surpasses the disadvantages (Velmurugan, 2013).

2.3 Status of Institutional Repositories in Zambia

MacDonald (2011) described IRs as libraries of digital objects accompanied by associated metadata originating from a single institution. These repositories serve to centralise, preserve, and provide access to the knowledge generated by HEIs. The advantages of IRs extend to HEIs by elevating their profiles, while scholars benefit from broader dissemination, increased utilization, and enhanced professional visibility of their scholarly research. Institutional Repositories serve as potent tools facilitating global access to the intellectual output of institutions, aiding in the preservation and maximization of access to research output on a global scale. The advent of open access has presented numerous opportunities for academic and research institutions, including establishing IRs. However, the development of IRs in most of African countries is still in its infancy, making most of African researchers unpopular and unrecognised in academic fraternity (Yusuf, Ifijeh, & Emmanuel, 2019; Phiri, 2018).

This trend persists, as Adam (2021) noted in an exploratory study on the status of IR implementation in African countries, revealed that the overall performance of IRs in Africa remains below average, with notable disparities among countries. South Africa, Kenya, Nigeria, Algeria, Sudan, and Egypt showed relatively more feasible prospects for global open access to research results through IRs. Despite significant efforts in support of open access implementation in Africa, the widespread implementation of IRs has been slow-paced, and the performance of implemented repositories still needs to improve, prompting further investigation into the situation on the continent. The above findings are similar to Phiri (2018) in a study entitled Research visibility in the Global South, where it was discovered that, out of the six public HEIs in Zambia, only Copperbelt University (CBU) and the University of Zambia (UNZA) had IRs implemented. These repositories were associated with a low count of digital objects, despite both HEIs graduating a relatively large number of postgraduate students then. The above findings could be attributed to the low

adoption of IRs in HEIs in Zambia due to lack of suitable ICT skill to manage IRs among the Librarian. Hence, a survey to assess the state of IRs in Zambia.

2.4 Metadata

Metadata is relatively new concept which denotes data about data. Metadata describes the attributes of information bearing object, document, data set, database, image, artifacts, collection etc. Metadata acts as a surrogate, representation of the content, context, structure, quality, place, condition and other characteristics of document for the purpose of representing the document to a potential user for discovery, evaluation, and for use, access, transfer, and citation (Dashrath, 2014).

There are three main types of metadata, and each type functions to make digital assets more and easily discoverable. These includes; structural metadata, descriptive metadata, and administrative metadata. Structural metadata, contains information necessary to record the internal structure of an item so that it can be rendered to the user in a sensible manner. It can also be said that structural metadata is data that indicates how a digital asset is organised. For example, pages in a book are organised to form chapters. Structural metadata also indicates whether a particular asset is part of a single collection or multiple collections and facilitates the navigation and presentation of information in an electronic resource. Example includes; page numbers Sections, chapters, indexes, and table of contents (Dashrath, 2014). Structural Metadata is primarily used to define relationships between digital objects and to facilitate effective presentation of digital objects, when rendered. Most IRs will present digital objects using container structures for organising related objects, and structural metadata plays a crucial role in associating digital objects to container structures. For IRs used in HEIs, the container structures will generally be associated with faculties, departments and other units that are associated with them (Phiri, 2020).

The administrative metadata provide information used to manage the resource e.g., when and how the resource was created and file type. Administrative metadata relates to the technical source of a digital asset. This type of metadata also relates to usage

rights and intellectual property, providing information such as the owner of an asset, where and how it can be used, and the duration a digital asset can be used under the current licence. Administrative metadata aims at providing information used for managing and processing digital objects, to facilitate the long-term preservation of digital objects and access control (International Standard organization, 2010).

The descriptive metadata is essential for discovering and identifying assets. It is information that describes the asset, such as the resource title, author, and relevant keywords. Descriptive metadata includes any information describing the resource that can be used for later identification and discovery (Dashrath, 2014). Descriptive metadata is the most robust of all the types of metadata its primary role is to enable effective discovery of digital objects. The descriptive metadata is generally encoded using a pre-defined metadata scheme and stored in a database management system. While textual IR digital object bitstreams that are born digital can easily be indexed using full-text features available in most IR platforms, some digital objects that are stored in different formats rely on descriptive metadata for their effective discovery.

2.5 Metadata Quality in Institutional Repositories.

Metadata creation stands as a fundamental activity for Libraries and information centres. The significance of quality metadata creation is on par with the care, preservation, display, and dissemination of collections. Adequate planning and allocation of resources are imperative for this ongoing and mission-critical endeavour. As Buca (2008); Tani, Candela and Castelli (2013) emphasised that, "more than the mere digitization of materials is needed to guarantee access." While digitization is a crucial step, it does not inherently make materials findable, understandable, or usable for the expanding audience of online users. However, when coupled with the creation of meticulously crafted metadata, digitization can significantly enhance end-user access, serving as the primary motivation behind the creation of digital resources.

In the pursuit of ensuring quality in metadata generation, Park (2010) observed that new metadata standards are evolving rapidly, and their introduction into various communities has prompted discussions on quality. However, the pace of this process varies among communities. The initial enthusiasm for metadata as a solution to information overload is often followed by the realisation that investments in quality are essential for meaningful improvements. Additionally, as communities aim to aggregate metadata using harvesting protocols like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), the absence of quality standards and measures becomes evident, as noted by (Suleman, 2012).

Alemneh (2009) evaluated the quality of metadata and identified the significance of high-quality metadata in aiding users in discovering information, especially in situations where users may need clarification about their exact requirements. The study recommends a comprehensive understanding of metadata records by incorporating detailed (micro) and overarching (macro) viewpoints. On a micro level, librarians should concentrate on the precise processes involved in producing quality metadata. In contrast, on a macro level, metadata should be considered within the larger framework of an information retrieval system. Building on Alemneh's findings, Ochoa (2014) also emphasised the significant association between metadata quality and effectiveness of digital libraries. The study highlighted that poor metadata quality compromises the discovery of digital library information objects, underscoring the critical role of metadata in information retrieval. Consequently, a stringent relationship between quality and adherence to bibliographic description practices must be maintained, as metadata quality profoundly influences the quality of services provided to users. A well-crafted metadata record enhances the value of a resource, if it is based on a thorough analysis of the resource itself.

Weagley, Gelches, and Park (2010) investigated interoperability and metadata quality within digital video repositories. Their study highlighted that poor metadata quality in Electronic Theses and Dissertations (ETDs) impedes achieving interoperability in digital libraries, both at the regional and national levels of digital library consortia.

The researchers underscored the need for robust metadata in an aggregated environment, emphasising that metadata should encompass information relevant to user queries for contents to be effectively discoverable. Dominic (2010) shares a perspective in line with Weagley, Gelches, and Park (2010), stressing that the limited visibility and commercial importance of Theses and Dissertations should not undermine their research value. Nevertheless, the quality of metadata instances stored in digital repositories is identified as a crucial concern affecting the effective operation and interoperability of these repositories. Dominic (2010) underscores the distinctive nature of ETDs as a specialised form of grey literature, highlighting their essential role in research. The significance of metadata quality in digital repositories becomes paramount for the effective functioning and interoperability of these repositories, directly impacting the primary purpose of a digital repository—to facilitate access to resources. To achieve successful interoperability in digital libraries, bibliographic records must share comparable fields and contain valid values within those fields. Attaining these properties requires standardised data entry practices and the adoption of common guidelines by record creators for describing content (Hafezi, et al 2010).

Various researchers, including Pattuelli, Contaxis, & Rossmann (2015); Dorodchi (2016); Bruce and Hillman (2004); Day and Masanès (2006), undertook evaluations of metadata record quality, considering factors such as accuracy, provenance, consistency, logical coherence, timeliness, accessibility, subject term specificity, and exhaustively. Despite the lack of a unanimous agreement on conceptual and operational definitions of metadata quality, the researchers agree on the paramount importance of metadata quality. They emphasise that assessing the appropriateness of metadata elements involves finding a balance between the specificity of the knowledge they convey and the associated cost of creating the descriptions. This highlights the ongoing significance of considering multiple dimensions when appraising the quality of metadata records. Therefore, HEIs must establish mechanisms to potentially facilitate ingesting high-quality ETD metadata. Many

issues highlighted in the literature are fundamentally linked to the comprehensiveness and correctness of metadata (Suleman, 2012).

As metadata plays a pivotal role in discovering and reusing digital objects. However, there is a growing concern about maintaining its quality. Information professionals have devised categories for the analysis and evaluation of metadata quality based on its ability to fulfil core bibliographic functions, including discovery, use, provenance, currency, authentication, and administration (Park, 2009). Park further emphasised that metadata instances should be exhaustive and comprehensive, enabling users to fully comprehend the purpose and content of the described resource without directly accessing it. The significance of metadata is also noted in distributed search, whose efficacy could be compromised when a connected repository contains predominantly low-quality metadata instances. Therefore, the utility of a digital repository is closely tied to the quality of the metadata describing its resources.

Ochoa (2009) delved into quality measurement criteria and identified three core benchmarks for practical metadata quality evaluation. These criteria, considered pivotal in assessing metadata quality, are completeness, accuracy, and consistency. As per Ochoa's evaluation, the completeness of a metadata record is determined by its total access capacity to individual local objects and its connection to the parent local collection(s). Park and Tosaka (2010) support Ochoa's findings by agreeing that completeness is not solely determined by filling numerous elements with values describing an object. Instead, they concur that completeness should be evaluated by considering the full access capacity to individual local objects and their connection to the parent local collection(s). This perspective underscores the idea that a comprehensive approach to accessibility and contextual linkage is crucial in measuring the completeness of metadata records.

Together, the above criteria furnish information professionals with a framework for evaluating current metadata and descriptive practices of HEIs in Zambia. As outlined above, the criteria align with the functional purpose of metadata in facilitating resource discovery and use. Therefore, in the context of this study, the assessment of

metadata comprehensiveness will involve verifying the ability to achieve full access to individual local objects and establishing connections to the parent local collection. This approach thoroughly evaluates the metadata's capacity to support effective resource discovery and usage in an aggregated environment.

2.6 Metadata Schemata

The diversity of information resources, formats, and disciplines necessitates acknowledging that there is no universally applicable metadata schema, controlled vocabulary, or data content standard exists. Therefore, HEIs must carefully select a suitable combination of metadata schemas and controlled vocabularies, including collection-specific thesauri from local pick lists (Pal, 2010). Additionally, adopting the most fitting cataloguing standards, which may involve local cataloguing guidelines based on established standards, is crucial for effectively describing and providing access to collections and other resources, as Baca (2008) notes. Moreover, the Australian National University[9] emphasises two central systems for metadata description, namely specialised and generalised schemes. Specialised schemes, while demanding more time and expertise, result in a more comprehensive description of the data, particularly when the data aligns with a developed schema. On the other hand, generalised schemes require less time and expertise but may yield less detailed descriptions, encompassing what are considered essential elements for describing any data. Given this perspective, many metadata schemes were developed to cater to the diverse information resources found in various formats and disciplines.

2.6.1 Dublin Core Metadata Scheme

The Dublin Core Metadata Element (DCME) Set is a standard for cross domain information resource description. It was developed to be simple and concise for describing Web-based documents. It is extensively used to describe digital materials such as composite media like Web pages, Image, Sound, Text and Video (Weagley, Gelches and Park, 2010).

---

[9] https://libguides.anu.edu.au/c.php?g=881167&p=6358464

The National Information Standards Organization (NISO) in 2009 highlighted the birth of Dublin Core Metadata scheme, whose element set originated from discussions at a 1995 workshop sponsored by the Online Computer Library Centre (OCLC) and the National Centre for Supercomputing Applications (NCSA). This scheme was designed as a general scheme, with primary goal of establishing a set of elements that authors could use to describe their web resources.[10] This was in response to the growing number of electronic resources and the inability of the library profession to catalogue all these resources. The Dublin Core aimed to define a small set of elements and simple rules that non-cataloguers could apply. This scheme had 13 original core elements that were later expanded to 15 elements, and they comprised Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. The Dublin Core was intentionally developed to be simple and concise, tailored explicitly for describing web-based documents (Weagley, Gelches, and Park 2010).

2.6.2 Darwin Core Metadata Scheme

Wieczorek, Bloom, Guralnick, and Stan (2012), observed that various metadata standards have been employed to facilitate the description and access of digital images of scientific specimens. The most popular among them is the Darwin Core (DwC) Metadata standards used for describing biological diversity data as it serves as an extension of the Dublin Core's metadata terms. It is a specialised scheme developed and being maintenance by the DwC group (Biodiversity Information Standards, 2014). The scheme is said to be the most used metadata standard to publish data about species occurrence worldwide due to its design which was primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information (Global Biodiversity Information Facility, 2019).

---

[10] https://www.oclc.org/content/dam/research/publications/library/2009/weibel-elis.pdf

2.6.3 Metadata Object Description Schema

Metadata Object Description Schema (MODS) is a schema for bibliographic element set that may be used for a variety of purposes, and particularly for library applications. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users.[11]

The Metadata Object Description Schema (MODS) and the Visual Resources Association's VRA Core are additional standard schemas that receive extensive support alongside Darwin Core (DwC) and provide more detailed descriptions than the simpler Dublin Core. Integrating these diverse metadata standards into the information retrieval environment has several advantages. One notable advantage is the ability to facilitate the batch ingestion of a broader range of materials without requiring additional transformational programs. Furthermore, this integration empowers each information retrieval community to choose a metadata standard that aligns best with their specific needs and content. This flexibility enhances the adaptability and utility of metadata standards within the broader context of information retrieval systems. As a result, diverse communities can effectively organise, describe, and access their digital resources according to their unique requirements and preferences.[12]

2.6.4 Electronic Thesis and Dissertations Metadata Scheme

The ETD-MS is a metadata standard that is based upon Dublin Core. It was developed by the Networked Digital Library of Theses and Dissertations (NDLTD). ETD-MS provides guidelines on the use of Dublin Core elements for the cataloguing of ETDs. "The standard consists of 21 elements:14 basic elements plus seven supplemental. Of the basic ETD elements, 13 correspond to basic Dublin Core elements (i.e., title, creator, subject, description, publisher, contributor, date, type, format, identifier, language, coverage, and rights). Two Dublin Core elements (i.e., relation and source) are not used by ETD-MS." (Park and Richard, 2011)

---

[11]https://www.loc.gov/standards/mods/

[12] https://ipt.gbif.org/manual/en/ipt/latest/darwin-core

The development and application of Electronic Theses and Dissertations (ETDs) metadata standards, along with the resulting quality, consistency, and interoperability of the metadata produced and exchanged, carry significant implications for discovering and preserving these distinct student works. As highlighted by Zhang (2014), the goal of interoperability is to construct cohesive services for users, bringing together technically diverse components managed by different organizations. The research emphasises that the effectiveness of ETD metadata standards relies on agreements to cooperate at three levels: technical, content, and organizational, ensuring a consistent interpretation of metadata values associated with specific elements.

Chapman, Reynold, and Shreeves (2009) observed a significant diversity of disciplines and content formats within Institutional Repositories (IRs). This diversity posed a challenge for employing a uniform metadata format and restricting only high-level controlled vocabularies across various IRs. The establishment of the ETD-MS in 1997 by the NDLTD became the de facto standard for describing ETDs in IRs. NDLTD re-designed the schema in 2003 to enhance its accuracy as a metadata standard tailored for ETDs. ETD-MS incorporates Dublin Core (DC) elements and includes additional elements explicitly designed for Theses and Dissertations (Potvin and Thompson, 2016). This schema marked a breakthrough in the organization and sharing of ETD metadata on a global scale, contributing to improved interoperability and standardization in the description of ETDs.

Park and Richard (2011) pointed out that ETD-MS is part of a broader initiative called the OAI PMH. This initiative served as a standard mechanism to gather metadata centrally and make it accessible to other institutions. The ETD-MS metadata standard is built upon unqualified Dublin Core, which employs 15 elements for item description. ETD-MS incorporates Dublin Core elements and introduces an additional element, "thesis." This extra element is accompanied by four qualifiers designed to capture information relevant to describing Electronic Theses and

Dissertations (ETDs): degree name, degree level, degree discipline, and degree grantor.[13] Consequently, ETD-MS comprises 21 elements: 14 essential elements and seven supplemental elements. Among the basic ETD elements, 13 align with basic Dublin Core elements (title, creator, subject, description, publisher, contributor, date, type, format, identifier, language, coverage, and rights). Notably, two Dublin Core elements, namely "relation" and "source," are not utilised in the ETD-MS.

Suleman (2004) pointed out that NDLTD played a significant role in establishing standards for information exchange across diverse sources within an interoperable framework. One of NDLTD's objectives was establishing authority control over personal and institutional names in ETDs within the NDLTD collection. The compilation of authoritative records, known as the Linked Authority File, was envisioned to consist of records generated and shared among participating institutions utilising the OAI protocol, with the associated costs shared. The adoption of the OAI protocol aims to enhance interoperability among different metadata formats. Suleman recommended that ETD-MS be adopted universally by HEIs because it is a flexible set of guidelines for encoding and sharing fundamental metadata related to ETDs. This adoption would facilitate the prudent organization and management of ETDs across IRs globally, with a specific emphasis on Zambia.

2.7 Generation of the Missing Metadata in Institutional Repositories

Riley's analysis (2017) of metadata generation practices within the cultural heritage sector spanning from the past to the present year revealed that descriptive metadata initially aided users in locating printed books, journals, manuscripts, or cultural objects. However, the process had no alternative but to rely on the human metadata creation. Elements like title, author, and publication date were transcribed manually after physically examining the items. Additional metadata, including value-added and interpretive information such as summaries or subjects, was similarly crafted and provided by experts. In contrast, the contemporary landscape reflects a shift where

---

[13] https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html#thesis.degree

descriptive, structural, and administrative metadata can now be generated not only by human experts but also through technical and automated processes. In addition, Dorbeva, Kim, and Ross (2013) noted in their study that recent technological advancements have showcased dependable approaches to generating metadata through automated processes, particularly in the context of born-digital information. Technical metadata serves as a notable example of this trend. Many file formats incorporate embedded technical information, intending to aid software in interpreting the content. Additionally, there is a precedent for software leveraging system-level details to append additional administrative information to digital files, such as the date of creation or the user ID logs into the system.

The Library of Congress Action Plan in 2005 acknowledged the necessity of employing automated methods to facilitate metadata generation, emphasising the positive impact on metadata quality. Tahmed and Mukhopadhyay (2023) also alludes that "automatic metadata generation is more efficient, less costly, and more consistent than human-oriented processing." While automated methods prove effective, they may encounter challenges, especially for metadata requiring human intellectual discretion. The authors suggested that the most effective outcomes in automatic metadata generation could be attained by integrating both human and automated methods. The observations above emphasise the library community's need to explore how automatic metadata generation can complement or serve as an alternative to traditional library metadata generation in IRs especially on regenerating incomplete metadata elements in IRs. Even though pure automatic metadata generation relies entirely on machine processing and is typically defined by its distinction from metadata generated by individuals many automatic metadata generation processes require human initiating the operations for machine manipulation of metadata to commence (Phiri, 2020; Greenberg, Spurgin & Crystal, 2005).

Dorbeva, Kim & Ross (2013) classified automatic metadata generation into three main categories: rule-based, neural networks-based and statistical modelling-based. Researchers often employ one of the mentioned methods, but these methods are not

mutually exclusive and can be combined. NYambe, Mulenga, Mufuzi, Ngoma, and Phiri, 2022; Azimjonov and Alikhanov (2018) engineered a rule-based system for metadata extraction from Portable Document Format (PDF). The researchers applied general layout rules, such as "title element is usually located on the upper portions of the first pages, and they are usually in the largest font sizes," this represented a common approach aimed at recognising and extracting specific text from a chosen format. Ongoing metadata extraction research often targets specific documents within the same file format. The document title is a commonly focused element in these extraction methods.

The remarkable expansion of digital resource repositories has provided a wealth of digital collections that serve as rich material for studying automatic metadata generation. In a study by Milena, Yunhyong, and Seamus (2013), the retrieval effectiveness (recall and precision) for queries against professionally and automatically generated metadata records was comparatively analysed. The metadata records represented the National Institute of Environmental Health Sciences web pages. The analysis of results from 10 queries regarding recall and precision suggested that professionally generated metadata records are not significantly superior in information retrieval effectiveness compared to automatically generated metadata records. Similarly, Mishra and Gupta (2017) conducted a comparable study, noting that vector analysis experiments leveraging document structure have been relatively successful. The research observed that using a Support Vector Machine (SVM) algorithm, coupled with 'word' and 'line' extraction, resulted in relatively high precision and recall ratios for metadata-based document retrieval.

2.8 Summary of Chapter 2

The chapter presented a review of the literature relevant to the study on IRs and quality metadata in HEIs in Zambia. IRs were first developed as an online solution for development, administration and disseminating of the scholarship for all HEIs globally. The repositories quickly transformed into platforms for publishing and showcasing the holdings of a given institution. The Added support for multimedia,

brought greater depth to repository collections. Since 2000, a number of repository platforms have been developed, each with their own set of benefits and technical criteria (Callicott, Scherer and Wesolek, 2016).

The literature review shown that The HEIs and other research entities globally reconsiders the production and usage of information resources in the digital format and they have been looking for ideal means for capturing and reusing the intellectual output for teaching and research. Little has been done on assessing metadata quality of ETDs in IRs particularly in Zambia as their implementation is still in infancy stage despite being viewed as playing vital role in transformation of academic knowledge management (Betz & Hall, 2015).

The IR platforms used varies from one HEI to another depending on its needs. However, there are two broader categories namely FOSS and proprietary software. Many countries have adopted FOSS mainly with desire to reduce the cost of acquiring software (Weagley, Gelches and Park, 2010). Critical in all these platforms is the assigning of adequate metadata to resources. Every digital object uses structural, administrative and descriptive metadata for it to be accessible online. Descriptive metadata is considered the most robust type of metadata, primarily enabling the effective discovery of digital objects. Its comprehensiveness is dependent upon the schema used as it is typically encoded using predefined metadata schemes and stored within database management systems.

The general schemes like Dublin Core have predefined elements that are encompassing all information resources online, in contrast to specialised schema like ETD-MS which is biased to ETDs. The other schemes like DwC for describing biological specimens and MODS which is used for a variety of purposes, particularly for library applications are some of schemes mostly used in IRs world over (Wieczorek, Bloom, Guralnick, and Stan, 2012).

Metadata generation practices from inception had no alternative but purely relied on the human metadata creation. All metadata were transcribed manually after technical reading technical. In contrast to contemporary landscape that reflects a shift where descriptive, structural, and administrative metadata could be generated not only by human experts but also through technical and automated processes (Riley, 2017). Dorbeva, Kim, and Ross (2013) noted in their study that recent technological advancements showcased dependable approaches to generate metadata through automated processes, particularly in the context of born-digital information. Azimjonov and Alikhanov (2018) engineered a rule-based system for metadata extraction from Portable Document Format (PDF). The researchers applied general layout rules, an approach that aimed at recognising and extracting specific text from a chosen format.

There were some gaps in the studies reviewed. The quality of ETD metadata is vital in successful implementation of national ETD portal. Many studies conducted indicated that downstream services are severely affected by poor ETD metadata originating from HEIs. However, extent has not been looked into from Zambian perspective due to lack of prior research in the same area. Hence, the need to conduct a study to addresses unanswered questions.

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

3.1 Overview

This chapter presents the research methods employed in the study. It furnishes information on the chosen research design, Target population and outlines sampling methodology. Additionally, the chapter expounds the rationale behind this selection and proceeds to describe the instruments utilised for data collection and delineates the procedures followed in conducting the study.

3.2 Research Design

The main objective of this study was to evaluate ETD metadata quality originating from HEIs in Zambia against ETD-MS, the de facto metadata scheme for ETDs developed by NDLTDs. Further, the study sought to establish the location of the missing ETD metadata elements from all IRs in Zambia.

The study's aims were translated into three research questions that needed to be addressed by collecting and analysing data from IRs from all HEIs in Zambia. These questions were:

i. Which Higher Education Institution implemented Institutional Repository that are functional and interoperable in Zambia?

ii. What is the relative quality of Electronic Theses and Dissertations Metadata in Higher Education Institutions of Zambia?

iii. What are the sources of missing Electronic Theses and Dissertations metadata in Higher Education Institutions in Zambia?

Given the comprehensive nature of research questions and the proposed methodology, a suitable research design adopted was Mixed-Methods Exploratory Study. This reflects the combination of qualitative and quantitative methods of data collection together with the exploratory nature of the investigation into the implementation of institutional repositories and the quality of ETD metadata in

Zambian HEIs. Kothari and Garg (2014), defined research design as arrangement of conditions for data collection and analysis in a manner that balances relevance to the research purpose with efficiency. It serves as a conceptual blueprint guiding the collection, measurement, and analysis of data, providing a structural framework for the research process.

3.2.1 Research Philosophy

Social research designs are modelled after many paradigms. The research philosophy underlying the above research questions involved a combination of realism, pluralism, objectivism, and constructivism, as they depended on the ontological and epistemological assumptions guiding the investigation into the implementation and assessment of institutional repositories and ETDs metadata quality in HEIs in Zambia (Park and Song, 2019).

3.2.1.1 Realism

This is ontological perspective that assume that IRs, ETDs, and their associated metadata have an objective existence independent of human perception. It suggests that there is a tangible reality to be studied regarding the implementation of IRs and the quality of ETDs metadata in HEIs.

3.2.1.2 Pluralism

Given the diverse nature of HEIs and their varied approaches to implementing IRs and managing ETDs metadata, the research design incorporated a pluralist ontological perspective. This perspective acknowledges the existence of multiple realities or perspectives, allowing for the recognition of different institutional contexts and practices

3.2.1.3 Objectivism

This epistemological perspective posits that knowledge is objective and can be acquired through systematic observation and empirical inquiry. It aligns with the quantitative aspect of the research design, where survey data and quantitative

measures are used to assess the presence of institutional repositories and the quality of ETDs metadata.

3.2.1.4 Constructivism

This epistemological perspective recognizes that knowledge is socially constructed and influenced by individual and collective interpretations. It is relevant to the qualitative phase of the research, particularly in interviews with stakeholders to understand their perspectives and experiences regarding institutional repositories and ETDs metadata quality.

3.2.2   Research Approach

A mixed-methods approach was used to leverage both qualitative and quantitative methods to address different aspects of above research questions. Qualitative methods provide depth and understanding, while quantitative methods offered wider breadth for generalizability. The combination of these approaches, helped to gain comprehensive understanding of the implementation of IRs and the quality of ETD metadata in Zambian HEIs (Kothari and Garg, 2014).

3.2.3   Validity and Reliability of the Study

Validity and reliability are crucial aspects of research methodology that ensure the trustworthiness and credibility of study findings.

When collecting data on HEIs that implemented IRs that are functional and interoperable in Zambia, content validity was considered by making sure that the survey questions and interview protocols were designed based on established definitions and criteria for functional and interoperable IRs. This ensured that the study accurately captured the intended constructs. The criterion Validity was also factored by making sure that survey responses and interview data was compared against objective criteria for functional and interoperable IRs that verified the accuracy of the findings. On reliability aspect, internal consistency was considered

and this was done by designing survey questions to measure the same underlying constructs consistently (Creswell, 2013).

During analysis of the relative quality of ETDs Metadata in Higher Education Institutions of Zambia, construct validity was adhered to by making sure that criteria used to assess metadata quality was be based on established standards and best practices in metadata management at the same ensuring that the study measured the intended constructs accurately. While cconcurrent validity was observed by ensuring that assessment criteria were compared against existing measures of metadata. Test-retest reliability was followed. A subset of ETDs was assessed by the same researcher at two different time points to assess the stability of the assessment criteria over time (Middleton, 2020).

While during identification of the sources missing Electronic Theses and Dissertations metadata in Higher Education Institutions in Zambia, content validity was adhered to by making sure that criteria used to identify sources missing metadata elements was based on established standards and guidelines for ETD metadata. This ensured that the study accurately captured the relevant dimensions of metadata completeness. While Inter-rater Reliability was adhered to by making sure that multiple researchers independently assessed the metadata completeness of the same subset of ETDs to assess the consistency and reliability of the findings (Middleton, 2020).

3.3 Identification of Higher Education Institutions with Functional and Interoperable Online Institutional Repositories in Zambia.

3.3.1 Study Setting

To meet the goals and objectives of this study, a survey of 61 HEIs in Zambia, comprising private and public institutions, was carried out (HEA, 2021).

3.3.2 Target Population

The target population in this study were 122 members of staff in charge of IRs in all HEIs i.e., 2 librarians per HEI.

3.3.3 Sampling Methods

Purposive sampling and total population sampling were used to sample all HEIs in Zambia to identify HEIs with IRs. Total population sampling was used on the premise that all HEIs must be included to have a valid result from HEIs that have adopted and implemented IRs for managing ETDs.

The purposive sampling was used to select two librarians managing IRs. This was in order to gather specific information from particular audiences involved in the management of IRs. Using this non-probability method, the irrelevant responses were filtered out that could have been collected had other methods been used.

The procedure for identifying HEIs that adopted and implemented IRs was done by consulting the HEA official website under registered HEIs' contact details, which was then used to obtain contact information regarding informants in every HEI.[14] After getting contact details for key informants, the structured questionnaire was sent online to all HEIs with IRs using Google Form

3.3.4 Measurement Instruments

3.3.4.1 Online Desk Research

This was done by checking the HEA website and consequently identifying points of contact that helped to provide information with regards to HEIs that adopted and implemented IRs.[15]

3.3.4.2 Questionnaires

---

[14] https://heaims.hea.org.zm/frontend/web/site/view-institution?id=20
[15] https://heaims.hea.org.zm/frontend/web/site/view-institution?id=20

A survey on the identified HEIs was carried out using an online structured questionnaire to get more information regarding the operations of IRs. The questionnaires were distributed using Google Forms. "Appendix C" shows the questionnaire used in the study.

3.4 Determining the Relative Quality of Electronic Theses and Dissertations Metadata in Higher Education Institutions in Zambia

3.4.1 Overview

This objective investigated the quality of ETD metadata records in IRs from all HEIs in Zambia. The study employed a completeness metric derived from (Bruce and Hillmann, 2004) to evaluate the quality of ETD metadata.

3.4.2 Target Population

The target population was all HEIs in Zambia. This was the case of wanting to evaluate all ETD metadata from HEIs in Zambia.

3.4.3 Data Collection

OAI-PMH was used to extract metadata from IRs and to collect the Dublin Core encoded ETD descriptive metadata records. The OAI-PMH protocol was developed to ensure interoperability standards to facilitate and encourage more extensive and efficient information transmission throughout the scientific community.

For HEIs that did not activate OAI-PMH, the researcher conducted online webinar training to sensitise HEI repository managers on the need to activate OAI-PMH and the advantages that would accrue to all institutions. This was done to motivate repository managers to activate a protocol that allows metadata sharing with local and global downstream portals. However, metadata quality analysis was not limited to IRs that activated OAI-PMH. For those that did not activate OAI-PMH, a manual approach was used. This involved downloading metadata records from all HEIs and then analysing the metadata to determine the completeness.

3.4.4 Procedure

The researcher accessed the URL for each IR and then accessed the OAI-PMH website to open the OAI-PMH tool interface. The researcher then selected the download XML interphase and put a valid Uniform Resource Locator (URL) for the repository, then clicked the download interface to download all metadata in XML format, e.g., University of Zambia IR. After ETD metadata were downloaded, the OpenRefine tool was used to convert XML files to Excel for easy analysis of the correctness and completeness of metadata. OpenRefine is a comprehensive tool designed to help to manage messy data sets with smoother and cleaner inconsistencies that commonly lead to errors.

For manually harvested ETD Metadata, the researcher visited all the identified IRs under Theses and Dissertations communities. This was because there were IRs that had not activated the OAI-PMH protocol, which necessitated manual content harvesting.

3.4.5   Sampling Method.

A stratified random sampling was used to select ETD metadata records from all IRs for document analysis due to the homogeneous nature of ETDs. Despite being produced from different schools and organised according to schools or faculty in IRs, ETDs share a standard structure and a layout prescribed in regulations postgraduate training handbooks in various HEIs.[16]

3.4.6   Sample Size

The sample size for metadata quality analysis was 56 ETD metadata records. This comprised of 8 ETD metadata records from all 7 HEIs that had functional IRs with ETDs resources. In qualitative research, sample adequacy focuses on the suitability of the sample's composition and size of a product. These are crucial aspects when assessing the validity of research (Vasileiou, Barnett & Thorpe, 2018). Therefore, in qualitative research, sample sizes are intentionally kept small to support an in-depth,

---

[16] https://graduate.unza.zm/graduate/regulations

case-oriented analysis. However, other scholars observed informational redundancy and information power as a principle guiding the determination of the sample size in qualitative research (Malterud, 2015). In view of homogeneous nature of ETDs, the sample size was arrived at after analysing 56 ETD metadata records from 7 HEIs.

3.5 Identification of Potential Sources of Missing Electronic Theses and Dissertations Metadata Originating from Higher Education Institutions in Zambia

3.5.1 Overview

This objective investigated the location of missing ETD-MS metadata elements from all IRs originating from HEIs in Zambia.

3.5.2 Target Population

Members of staff in charge of research in all HEIs, ETD manuscripts and policy documents that regulate postgraduate training from all HEIs in Zambia.

3.5.3 Sampling Methods

Stratified random sampling was used to select ETDs manuscripts for content analysis due to their homogeneity. A total population sampling was used to sample all policy documents and research coordinators from all HEIs to understand the outline and prescribed format in which ETDs must be presented in every HEI.

3.5.4 Sample size

In this study, the saturation principle was used to determine the sample size of ETDs for inclusion in the analysis due to their homogeneous nature, which are all Theses/ dissertations by type and share one standard structure and layout accepted in every HEI. Thus, saturation was confirmed after no new data was gathered after analysing 8 ETD manuscripts (PDF documents) per HEI. Therefore, 56 ETDs (8 Per HEI) and 7 policy documents i.e. (1 Per HEI) and 7 research coordinators (1 per HEI). The entire population for policy documents was sampled because there is only one document per HEI.

51

3.5.5   Data Collection

3.5.5.1   Online Desk Research

ETD manuscripts and policy documents that regulate how ETDs must be presented with regard to outline and format were accessed from HEIs websites. ETD manuscripts and policy documents were downloaded separately and analysed to identify the location of missing ETD-MS metadata. For those HEIs who did not upload their policy documents, key informants were contacted and requested for files that guides the postgraduate programs.

3.5.5.2   Interviews

For HEIs that did not produce a comprehensive document explaining the standard outline that regulates how ETDs must be presented, staff in charge of research was identified and contacted for interviews. Appendix B shows the interview guide used in this study.

3.6   Data Analysis

3.6.1   Identification of Higher Education Institutions that have Functional and Interoperable Online Institutional Repositories in Zambia

Data collected using structured questionnaires and the OAI-PMH Validator & data extractor tool was analysed quantitatively using a spreadsheet application, Google Sheets, and qualitatively using content analysis. Tables and graphs were used to present the analysed data.

3.6.2   Determining the Relative Quality of Electronic Theses and Dissertations metadata in Higher Education Institutions in Zambia

Data collected on this objective was analysed quantitatively using Google sheets and qualitatively using content analysis of randomly sampled ETD manuscripts from HEIs with functional IRs. Google sheets were used to present the analysed data in tables and figures.

3.6.3   To Identify Potential Sources of Missing Electronic Theses and Dissertations Metadata Originating From Higher Education Institutions in Zambia

The document analysis was used to analyse policy documents against the prescribed standard outline and format. Content analysis was used to analyse ETD manuscripts to ascertain the location of ETD-MS metadata from preliminary pages' sections. This data was quantitatively analysed into figures and tables using Google Sheets.

3.7   Ethical Considerations

Cohen (2016) explained the role of ethics as a guide to dos and don'ts that the researcher must observe during the research process to respect the rights and privacy of the respondent. The consent to conduct this study was obtained from the University of Zambia, Humanities and Social Sciences Research Ethics Committee (HSSREC). Appendix F: shows the University of Zambia Humanities and Social Sciences Research Ethics Committee Ethical Clearance Approval. In all phases of this research, the researcher observed confidentiality in the process of conducting research, thereby ensuring that all information collected was treated with utmost privacy. Anonymity was taken into account by not capturing the names of respondents in the presentation of results. The researchers did not use embarrassing questions or threatening statements when collecting data. The study required informed consent from respondents. In addition, the data collected was used purely for academic purposes and not for any other gains.

3.8   Summary of Chapter 3

The research adopted a mixed methods design which was used to survey 61 HEIs registered with HEA. Data on HEIs with functional and interoperable online IRs was collected using structured questionnaires and semi structured phone interviews. This data was analysed quantitatively using Google Sheet and Microsoft excel.

Data on determining the relative quality of ETD metadata originating from HEIs in Zambia was collected using OAI PMH tool and manual downloading. This data was analysed using Google sheet and content analysis. While data on Investigation of the

sources of missing metadata elements for the automated generated using NLP techniques was collected using Content analysis and Interviews. The data analysis was done using Google Sheet for quantitative content analysis. The next chapter presents the findings of the study

CHAPTER FOUR: RESULTS

4.1 Overview

This chapter presents the findings of a study that investigated Quality of ETD metadata generated from HEs in Zambia. The chapter is organised into several sections, examination of HEIs in Zambia with functional and interoperable online Institutional Repositories (IRs), an assessment of the quality of ETD metadata in these institutions, and determining the location missing ETD-MS metadata.

4.2 Higher Education Institutions in Zambia that have Functional and Interoperable Online Institutional Repositories in Zambia

Zambia has 61 HEIs registered with the Higher Educational Authority (HEA, 2021). Out of 61 Universities, we had 48 responses. The analysis of responses only included completed ones. Consequently, all incomplete responses were excluded from the analyses. Multiple responses from the same institutions were equally removed and only remained with unique ones per HEI. Therefore, this analysis was based on unique and complete responses only.

4.2.1   Distribution of Higher Educational Institutions in Zambia

At the time of data collection, it was found that on had Zambia only had 9 public functional universities and 52 private universities. Results indicated that 36 Out of 61 HEIs offer postgraduate programs, as shown in Table 1

Table 1 Distribution of Higher Educational Institutions in Zambia

Type of HEI	Count	HEIs offering Postgraduate programmes
Number of Public Universities	9	8
Number of Private Universities	52	28

4.2.2 Institutional Repositories in Zambia

To ascertain HEIs with functional and interoperable IRs, a survey in 61 HEIs was conducted using questionnaire and interviews. The research found that 10 (16.39%) HEIs in Zambia had implemented IRs, and 51 (83.6 %) of HEIs were still managing their ETDs in analogue format as shown in figure 1.
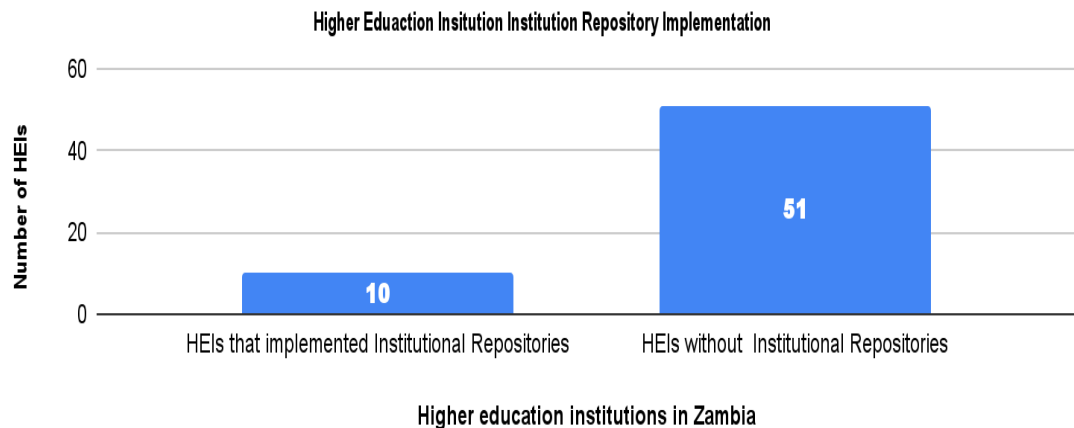
Figure 1 Institutional Repositories Implementation in Zambia

4.2.2.1 Functional Status of Institutional Repositories in Zambia

Out of 10 that implemented IRs in Zambia, 2 (20%) were found to be non-functional due to technical faults during data collection, 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 figure 2

Functional status of Institutional Repositories In Higher Educational Institutions
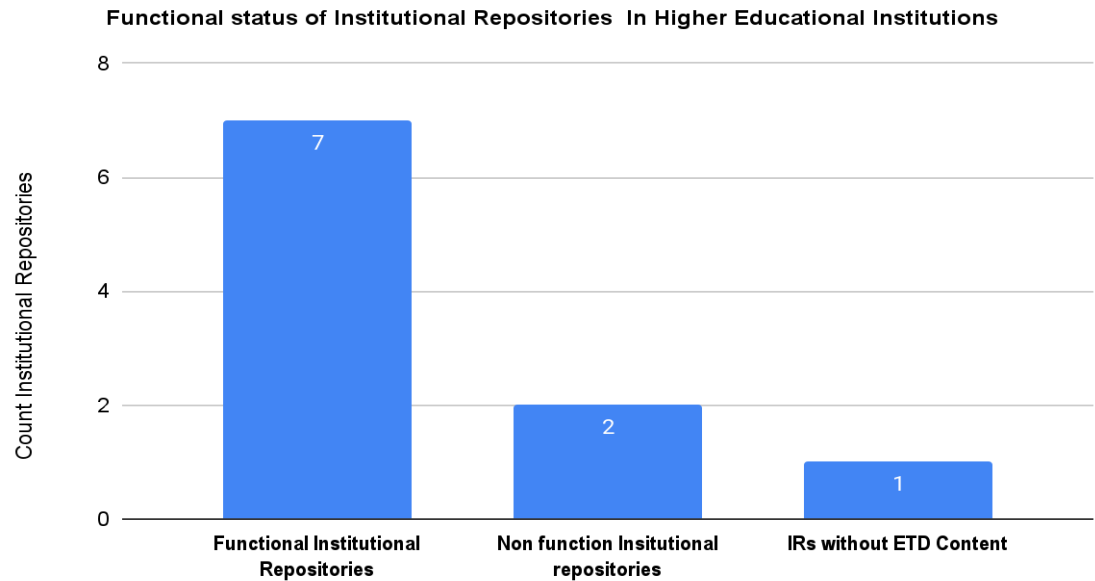
Figure 2 Functional Institutional Repositories in Zambia

4.2.2.2 Platforms (Software) used in Institutional Repositories in Zambia

All the 10 Universities that adopted and implemented the IRs were found to be using the DSpace platform (software) to implement IRs with Dublin core metadata schema, as shown in Table 2.

Table 2 Platforms (Software) used in Institutional Repositories in Zambia

Higher Education Institutions	Fedora Commons	EPrints	DSpace	Other
Nkwame Nkruma University	X	X	✔	X
University Of Zambia	X	X	✔	X
Chalimbana University	X	X	✔	X
University Of Lusaka	X	X	✔	X
Cavendish University	X	X	✔	X
Copperbelt University	X	X	✔	X
Lusaka Medical Apex University	X	X	✔	X
Mulungushi University	X	X	✔	X
ICT University	X	X	✔	X
ZCAS University	X	X	✔	X

4.2.2.3 Metadata Schema Used in Institutional Repositories in Zambia

All the IRs were assessed to determine the metadata schemes used to describe ETDs, the results showed that all IRs used the Dublin Core metadata scheme, as shown in Table 3

Table 3 Metadata schema used in Institutional Repositories in Zambia

Schemes	Number of HEI using the scheme
Dublin Core	10
DwC	0
ETD-MS	0
MODS	0

4.2.2.4 Interoperability Status of Institutional Repositories in Zambia

Among the functional IRs, it was found that 8 IRs representing 80% did not activate the OAIP-MH, and only 2 representing 20 % activated the OAIP-MH used to facilitate harvesting of metadata by the ETD portals, as shown in figure 3.
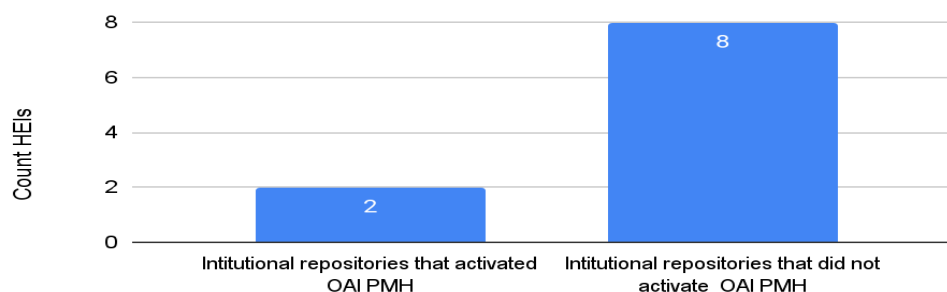
Figure 3 Open Archive Initiative Protocol for Metadata Harvester Activation and Configuration in Zambia.

4.3 Relative Quality of Electronic Theses and Dissertations Metadata in Higher Education Institutions in Zambia

Metadata from all IRs were harvested and analysed to ascertain their quality using completeness metrics.

4.3.1 The University of Zambia Institutional Repository

In a quest to analyse the ETD metadata quality at UNZA IR using ETD-MS metadata schema as de facto standard, the results shows that UNZA IR had over 10,000 records of electronic resources on (23rd February 2022), which were harvested. After data cleaning, we found that UNZA IR contained only 4003 records of ETDs and used 9 Dublin Core metadata elements to describe ETDs. The elements used include dc. title, dc. creator, dc. subject, dc. description, dc. publisher, dc. date, dc.type, dc. identifier and dc. language.

Out of the 4003 ETD records, the following ETD-MS elements were found missing: 350 (8.74%) publisher elements, 845 (21.11%) creator elements, 814 (20.33%) subject elements, 651 (16.26 %) type elements, 4003 (100%) contributors elements, 4003 (100%) thesis. degree elements, 12 (0.3) language elements, 4003 (100%) coverage elements, and 910 (22.7 %) records had description elements missing, as shown in Figure 4.

Missing ETD-ms Schema Elements at the University of Zambia

ETD-MS Schema Elements	Missing ETD-MS metadata in Percentages (%)
dc. title	0.00
dc.subject	20.33
dc.description	22.73
dc.publisher	8.74
Contributor	100.00
creator	21.11
dc.date	0.00
dc.type	16.26
dc.format	0.00
dc.identifier	0.00
dc.language	0.30
dc.coverage	100.00
dc.rights	0.00
thesis.degree	100.00
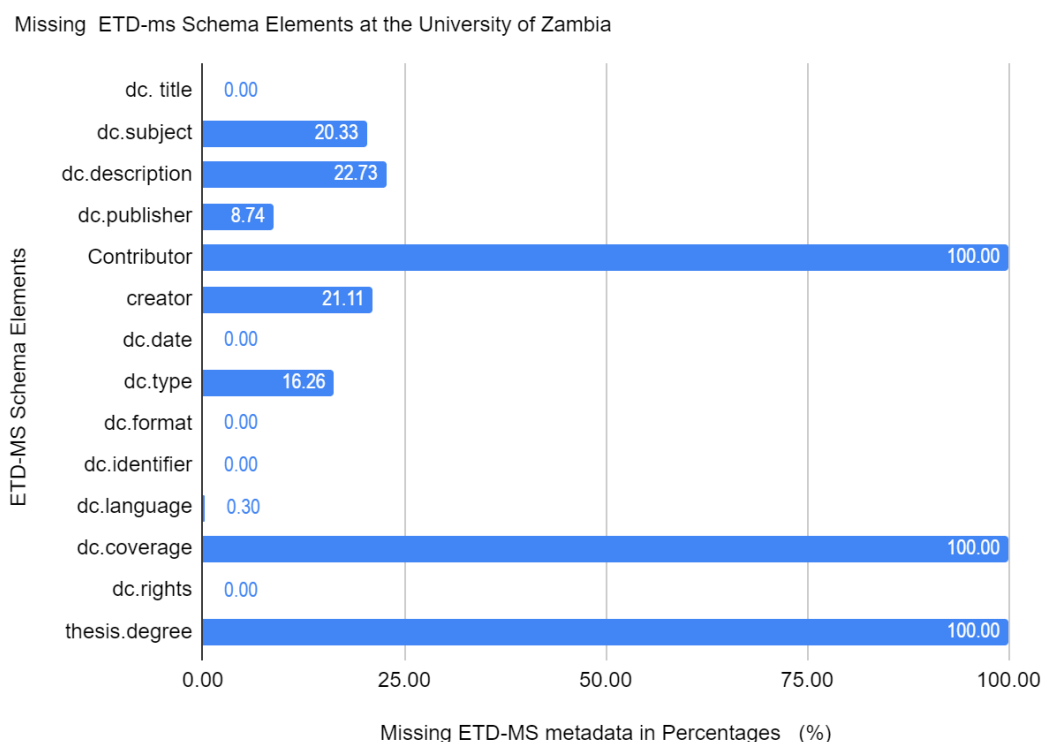
Figure 4 Missing ETD-MS Metadata Elements for University of Zambia Institutional Repository

4.3.2 University of Lusaka Institutional Repository

The total number of bibliographic records for ETDs in the UNILUS IR by 28th June 2022 was 37. ETD 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 IR uses 7 Dublin Core elements that comprise dc. creator, dc.date. dc.identifier, dc.language, dc.subject, dc.title, and dc.type.

UNILUS ETD Metadata records 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 all the 37 bibliographic records in the repository had six (6) ETD-MS metadata elements missing. The missing elements included dc. contributor, dc. format, thesis. degree, and dc. coverage, as shown in Figure 5

Missing ETD Metadata from University of Lusaka
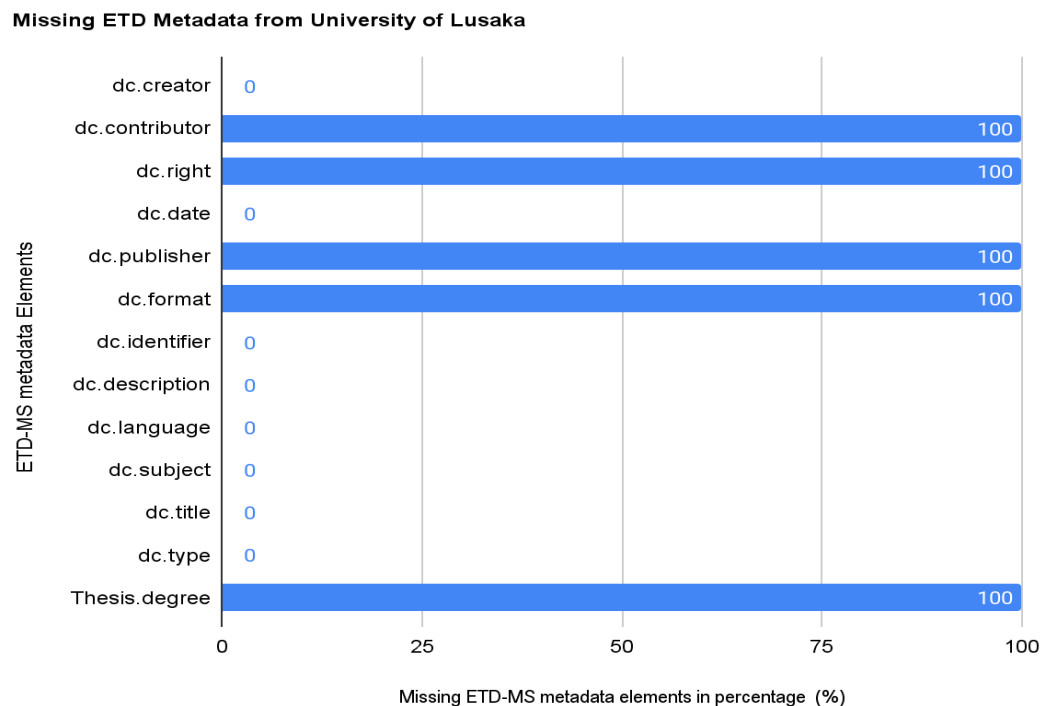
Figure 5 Missing ETD-MS Metadata used at University Lusaka

4.3.3 Cavendish University Institutional Repository

The Electronic Theses and Dissertations metadata from Cavendish University (CUZ) were harvested and analysed to identify metadata elements that are used to describe ETDs in the repository. The results indicated that there were 881 records of ETDs ingested in the CUZ IR by 28th June 2022. An investigation into metadata elements indicated that CUZ IR uses 8 metadata elements to describe ETDs. These 8 Dublin Core elements, comprised; dc. date, dc. identifier, dc. language, dc. description, dc. subject, dc.title, dc.type, and dc.creator.

The ETD metadata records from CUZ IR were further analysed against completeness and compliance to ETD-MS, the de facto standard for ETD Metadata. The study revealed that CUZ IR had 4 missing ETD-MS elements in all 881 records, and the missing elements included 881 (100%) dc. contributor elements, 881(100%) dc. rights, 881 (100%) thesis. degree and 881 (100%) dc. format as shown in figure 6
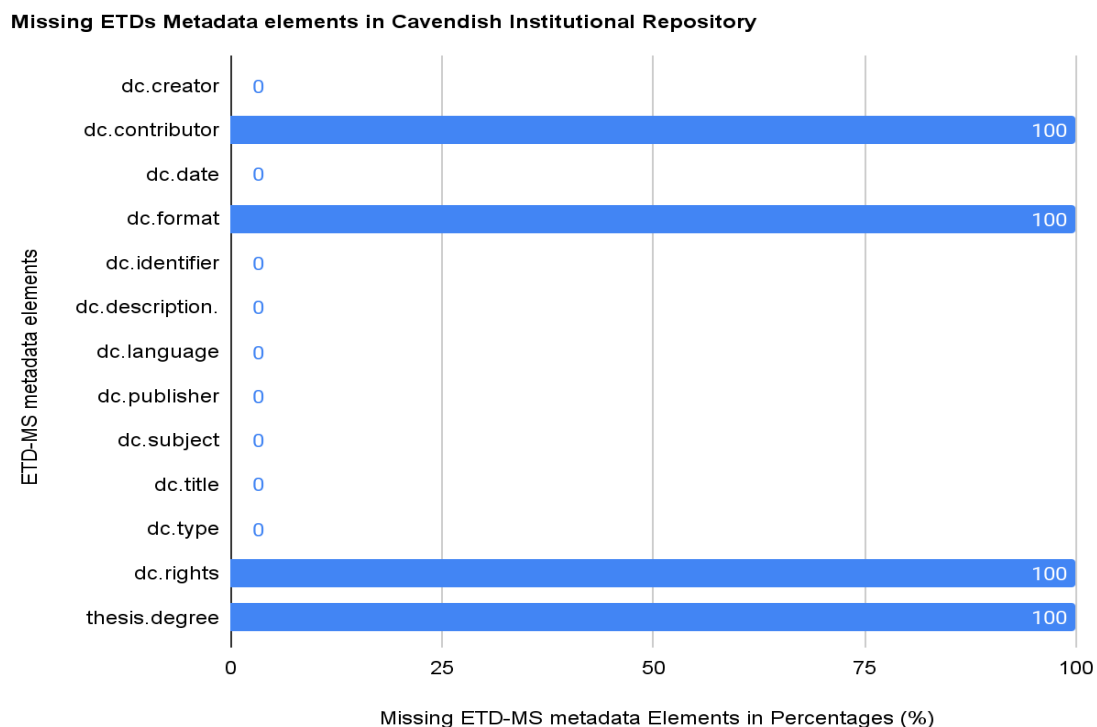
Figure 6 Missing ETD-MS Elements From Cavendish University Institutional Repository

4.3.4   Mulungushi University Institutional Repository

On 15th September 2022, MU IR had a total of 21 ETDs. When ETD metadata from MU were analysed to determine the metadata elements used, results showed that MU IR uses 9 Dublin Core elements to describe ETDs. These elements include dc. creator, dc.date accessioned, dc.identifier, dc.language, dc.description, dc.subject, dc.title, and dc.type.

In a quest to analyse the ETD metadata compliance against ETD-MS, the de facto metadata schema, the researcher discovered that MU IR had 5 ETD-MS elements missing. The missing elements included dc. contributor, dc. coverage, dc. rights, thesis. degree, and dc. format, as shown in Figure 7
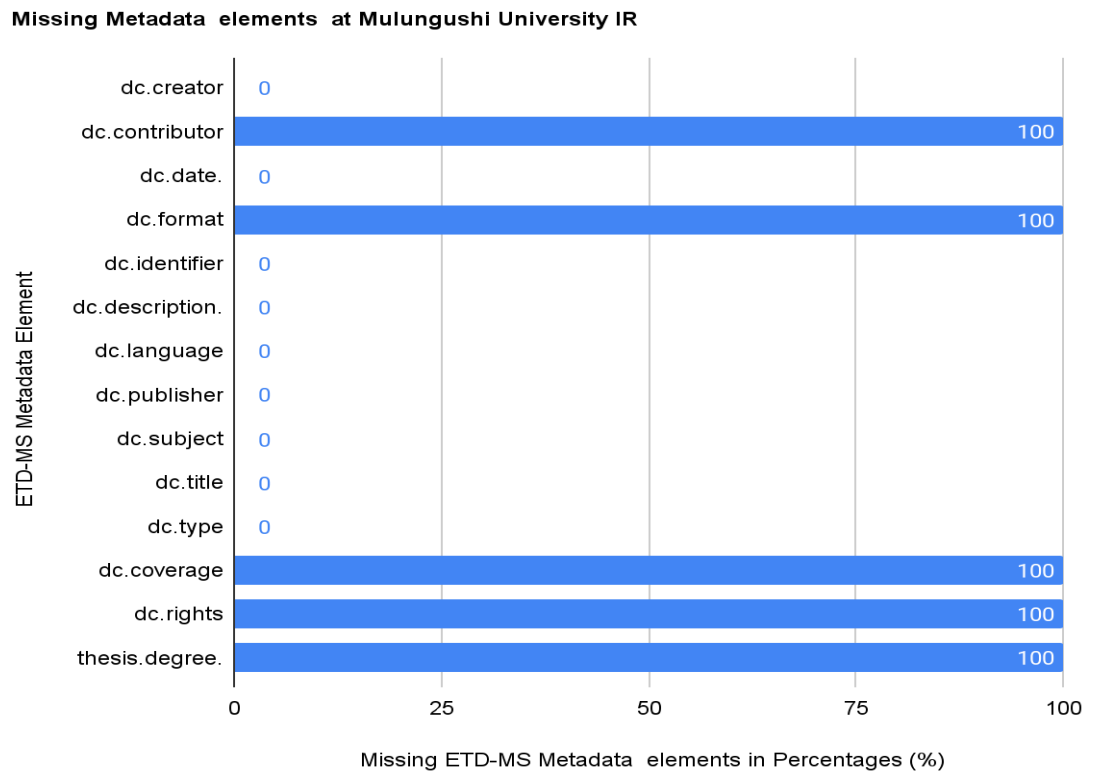
Figure 7 Missing ETD-MS Elements From Mulungushi University

4.3.5 ZCAS University Institutional Repository

ZCAS IR had a total number of 39 ETDs on 26 September 2022. The ETD metadata records from ZCAS University were harvested and analysed to ascertain the metadata elements used in describing ETDs. The findings showed that ZCAS IR uses 8 Dublin Core elements when creating bibliographic records for ETDs. The elements used include dc.creator, dc.date, and dc.identifier. dc.language. dc.description, dc.subject, dc. title and dc. type.

The ETD metadata records from ZCAS IR were also subjected to quality analysis using completeness metrics and their compliance against ETD-MS, the de facto metadata schema for ETDs developed by NDLTD. The results showed that ZCAS IR had five 5 missing elements in all 39 records of ETDs. The missing elements included dc. contributor, dc. coverage, dc. rights, thesis.degree, and dc.format, as shown in figure 8
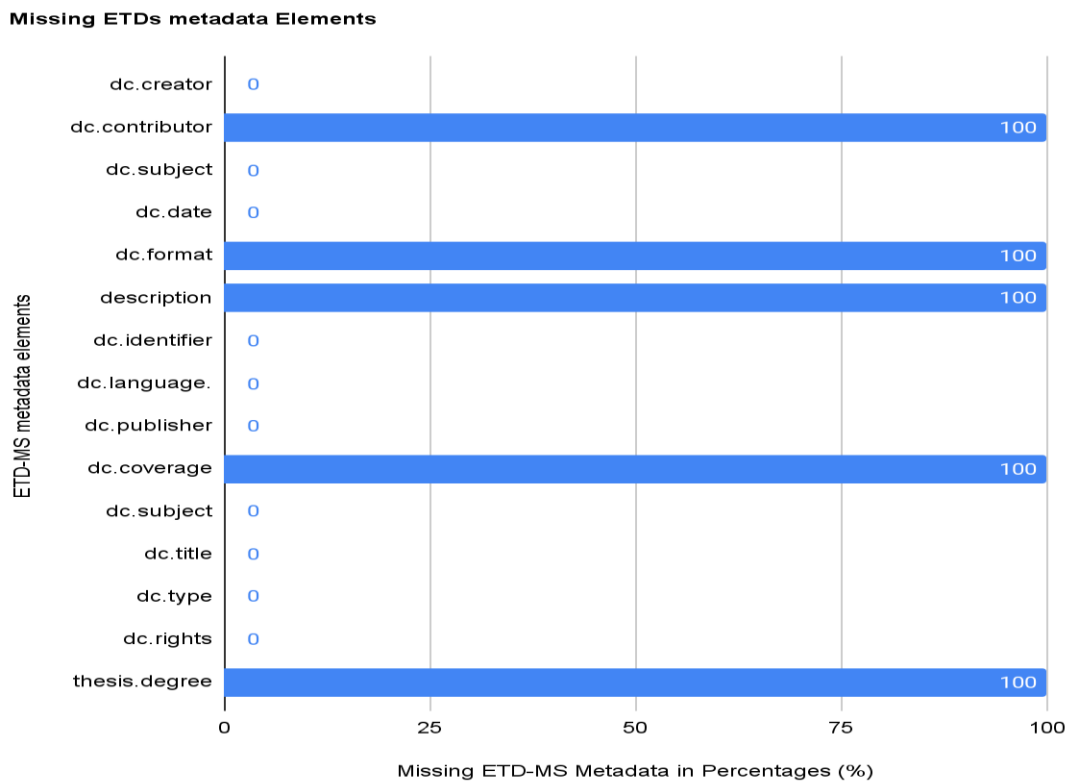
Figure 8 Missing ETD-MS Elements From ZCAS University

4.3.6 Chalimbana University Institutional Repository

Chalimbana University (CHAU) IR had 10 ETDs by 25th October, 2022. All ETD metadata were harvested and analysed to ascertain the metadata elements used for bibliographic description. The findings showed that CHAU IR used 8 Dublin Core metadata elements to describe ETDs. The elements included: dc. author, dc. date, dc. identifier, dc.language, dc.description, dc. subject, dc. title and dc.type.

Chalimbana University ETD metadata records were also subjected to quality analysis using completeness metrics and compliance to ETD-MS, the de facto metadata schema for ETDs. The research discovered that in all 8 records sampled, CHAU IR had 6 missing ETD-MS elements. The missing elements included dc. contributor, dc. coverage, dc.rights, and thesis. degree, dc.publisher and dc.format as shown in figure 9

Missing ETD-MS metadata elements at Chalibana University
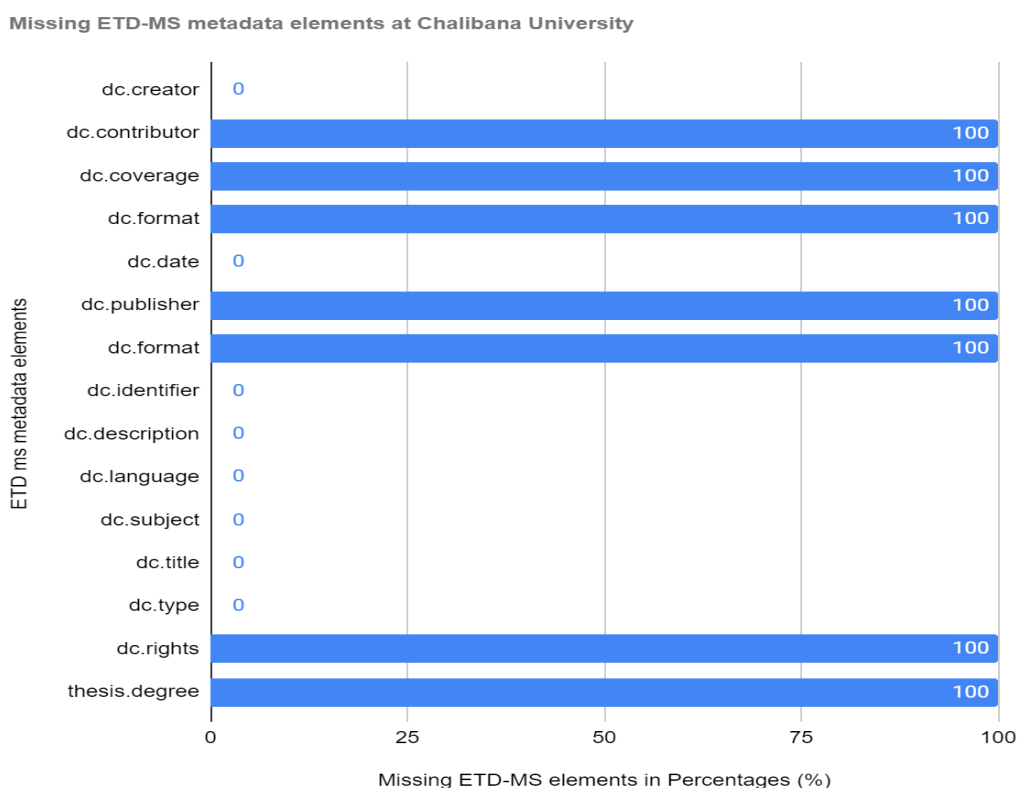
Figure 9 Missing ETD-MS Elements From Chalimbana University

4.3.7 Lusaka Apex Medical University Institutional Repository

LAMU IRs by 28 December 2022 had 13 ETD records. The ETD metadata records from the Lusaka Apex Medical University (LAMU) were harvested and analysed to understand metadata elements used in describing ETDs. The research found that LAMU IR uses 9 metadata elements when creating bibliographic records of ETDs. The elements used include dc.creator, dc.date, dc.identifier, dc.language, dc.description, dc.subject, dc.title, dc.publisher, and dc.type.

The ETD metadata from the LAMU IR were further analysed against completeness and compliance to ETD-MS, the de facto standard for ETDs. The study reviewed that LAMU IR had 4 missing ETD-MS elements, and these include dc.rights, dc.contributor, dc.format, and theses.degree, as shown in figure 10

Figure 10 Missing ETD-MS Metadata elements From Lusaka Apex Medical University

4.4 Sources of Missing ETD-MS Metadata Elements for Electronic Theses and Dissertations From Higher Education Institutions in Zambia

Unstructured interviews with research coordinators and content analysis of regulation for postgraduate training handbook and ETD manuscripts was conducted from IRs in all HEIs Zambia.

4.4.1 University of Zambia Institutional Repository

Results showed that the UNZA had the postgraduate handbook which was accessible on the UNZA website.[17] A content analysis on ETDs manuscripts and Regulations for Postgraduate Training handbook showed that section 26 of the postgraduate handbook contained the guidelines for formatting ETDs. The results clearly showed the sections in the manuscripts where students must put all information related to ETD metadata, as shown in Table 4

Table 4 Location of ETD Metadata From the University of Zambia Institutional Repository

DRGS SECTION	MANUSCRIPT SECTION	ETD-MS elements
Section 26	Title Page	dc.title
		dc.type
		dc.creator
		dc.publisher
		dc.date
		thesis.degree
		thesis.degree.department
		thesis.degree.discipline
	Copyright Declaration	dc.rights
	Certificate of Approval	dc.creator
	Abstract	dc.description
		dc.subject
	Acknowledgements	dc.contributor

4.4.2 Mulungushi University Institutional Repository

A survey at Mulungushi Directorate of Research and Graduate Studies (MU DRGS) to ascertain the availability of a postgraduate handbook that guides students on the structure and format of the ETDs, the results indicated that MU had the Regulations for Postgraduate Training handbook. A document analysis on the postgraduate handbook indicated that section 11.0 of the DRGS contained information about the acceptable structure and the format of the ETDs.

A survey on ETDs at the MU IR revealed that the organization of ETDs was according to disciplines. A content analysis on the ETD manuscripts indicates the location of Missing ETD-MS metadata as shown in Table 5.

Table 5 Location of ETD Metadata at the Mulungushi University Institutional Repository

MU DRGS Section 11.0	Thesis preliminary pages/section	ETD-MS metadata elements
	Title page	dc.creator
		theses.degree
		dc.publisher
		dc.date.
		dc.contributor
	Copyright Declaration	dc.right
	Certificate of Approval	dc.publisher
		dc.contributor
	Declaration	Creator's name
	Supervisor's Recommendation	dc.contributor
	Acknowledgement	Supervisor names
	Abstract	dc.description
		dc.subject

4.4.3   University of Lusaka Institutional Repository

An assessment at the University of Lusaka postgraduate section indicated that UNILUS has explicit document with a title; Research Proposals and Dissertations/Theses guidelines. An analysis of the postgraduate document shows that section 3.3 of the file contains information on the structure and format students must adhere to as they prepare their ETDs.

A study on ETDs at the UNILUS revealed that ETDs were organised in a broader community called ETDs Collection with a sub-community called Master's Theses and Ph.D./DBA Thesis. A document content analysis was conducted to determine the location of metadata on the ETDs' preliminary pages, the results are shown in Table 6.

Table 6 Location of Missing ETD Metadata at the University of Lusaka

Section 3.3 of Guidelines for Research Proposal and Dissertation/Thesis	Thesis preliminary pages / section	ETD-MS metadata elements
	Title page	dc.title
		dc.creator
		thesis.degree
		dc.type
		dc.contributor
		dc.publisher
		dc.date
	Copyright Declaration	dc.rights
	Certificate of Approval	dc.contributor
	Declaration	dc.creator
	Acknowledgement	dc.contributor
	Abstract	dc.description dc.subject

4.4.4 Lusaka Apex Medical University Institutional Repository

An inquiry on the presence of a postgraduate handbook at the LAMU shows that there was yet to be a completed document, as work was reportedly underway to develop one.

A survey on ETDs at LAMU manuscripts showed that the organization of ETDs was according to schools. A document analysis was conducted to determine the location of missing ETD-MS metadata from preliminary pages, the results are shown in Table 7.

Table 7 Location of Missing ETD Metadata from Lusaka Apex Medical University Institutional Repository

Thesis preliminary pages/section	ETD-MS metadata elements
Title page	dc.title
	dc.creator
	thesis.degree
	dc.publisher
	dc.contributor
	dc.date
Copyright Declaration	dc.right
Supervisor's declaration	dc.contributor
Acknowledgement	dc.contributor
Abstract	dc.subject dc.description

4.4.5 Cavendish University Institutional Repository

An inquiry on the availability of a postgraduate handbook at the CUZ indicated the unavailability of the document on guidelines for postgraduate studies at the institution, as work was still underway to develop it. A survey on ETDs at Cavendish University shows that the organisation of ETDs was according to schools, which are four in total. Using Stratified random sampling to select ETDs from UCZ IR, a total

sample of 8 ETDs was selected on which document content analysis was conducted to determine the location of metadata on the ETDs' preliminary pages. Results are shown in table 8.

Table 8 Location of Missing ETD Metadata in Cavendish University

Theses preliminary pages / section	ETD-MS metadata elements
Title page	dc.title
	dc.creator
	thesis.degree.
	dc.publisher
	dc.date
	dc.contributor
Copyright Declaration	dc.right
Declaration	dc.creator
Acknowledgement	dc.contributor
Abstract	dc.subject
	dc.description

4.4.6 Chalimbana University Institutional Repository

A survey on the availability of a postgraduate handbook at the CHAU shows no document on guidelines for postgraduate studies at the institution, as it was underworked. At the same time, a study on CHAU ETD manuscripts indicated that ETDs were organised according to a broader community called thesis and dissertation. A stratified random sampling of 8 ETDs out of 10 total populations of ETDs from the CHAU IR was done, on which document content analysis was conducted to determine the location of missing metadata on the ETDs' preliminary pages. Results are shown in Table 9.

Table 9 Location of the Missing ETD Metadata From Chalimbana University

Theses preliminary pages/sections	ETD-MS metadata elements
Title page	dc.title
	dc.creator
	thesis.degree.discipline
	thesis.degree.level/Name
	dc.publisher
	dc.date
Copyright Declaration	dc.right
Declaration	dc.creator
Approval	dc.contributor
Acknowledgement	dc.contributor
Abstract	dc.subject
Declaration	dc.contributor

4.4.7 ZCAS University Institutional Repository

A survey on the availability of a postgraduate handbook at the ZCAS shows that there is no document on guidelines for postgraduate studies at the ZCAS, as a committee was being constituted to investigate it. At the same time, a study at ZCAS IR on ETD manuscripts indicated that the organization of ETDs was according to a broader community called Thesis and Dissertation. A stratified random sampling of 8 ETDs from the ZCAS University repository was done, on which document content analysis was conducted to determine the location of missing metadata elements on the ETDs. The results are summarised in Table 10

Table 10   Location of the Missing ETDs Metadata From ZCAS University
Institutional Repository

Thesis preliminary pages / section	ETD-MS metadata elements
Title page	dc.title
	dc.creator
	thesis.degree.
	dc.publisher
	dc.contributor
	dc.date
Copyright Declaration	dc.right
Declaration	dc.creator
Acknowledgement	dc.contributor
Abstract	dc.description dc.subject

4.5 Summary of Chapter 4

This chapter presented the results of the study on institutional repositories and metadata generations in higher educations in Zambia. The results have been presented in line with the three objectives of the study and the results shown that Zambia has 61 HEIs registered with the Higher Educational Authority (HEA, 2021). Out of 61 Universities, we had 48 responses. At the time of data collection, it was found that Zambia only had 9 public functional universities and 52 private Universities. Results indicated that 36 Out of 61 HEIs offer postgraduate programs.

When ascertaining HEIs with functional and interoperable IRs, results found that 10 (16.39%) HEIs in Zambia had implemented IRs, and 51 (83.6 %) of HEIs were still managing their ETDs in analogue format. Out of 10 that implemented IRs in Zambia, 2 (20%) were found to be non-functional due to technical faults during data collection, and 1 (2%) did not have ETDs content despite offering postgraduate programs at the time of data collection and analysis of results. When analysed against the platform used, all the 10 HEIs that implemented the IRs were found to be using the DSpace platform with Dublin core metadata schema. An assessment of interoperability among the functional IRs found that 8 IRs representing 80% did not

activate the OAIP-MH, and only 2 representing 20 % activated the OAIP-MH used to facilitate harvesting of metadata by the ETD portals.

Metadata from all IRs were harvested and analysed to ascertain their quality using completeness metrics, the results showed that all Institutional repositories were not ETD-MS compliant to de facto ETD schema developed by NDLTD. A further analysis with desire of locate the missing ETD-MS elements, policy documents and ETDs manuscripts analysed indicated diverse sources of missing metadata elements from various IRs. For instance, while dc.contributor details were found on the title page of most manuscripts for most HEIs, these details were not available in manuscripts for some HEIs.

CHAPTER FIVE: DISCUSSION AND INTERPRETATION OF FINDINGS

5.1.　Overview

This chapter discusses the findings presented in Chapter 4. The discussion highlights the significant aspects based on the following research objectives:

i. To identify HEIs that have functional and interoperable online IRs in Zambia.

ii. To determine the relative quality of ETD metadata in HEIs in Zambia.

iii. To Identify potential sources of missing ETD metadata originating from HEIs in Zambia

5.2.　Higher Education Institutions and Functional and Interoperable Online Institutional Repositories in Zambia

According to the study's findings in section 4.2 of chapter 4, most HEIs are yet to implement IRs in Zambia as only 10 out of 61 HEIs have adopted and implemented IRs in Zambia. These results are consistent with earlier studies by Phiri (2018) and Adam (2021), whose results revealed that the widespread implementation of IRs still needed to be enhanced, and the performance of the implemented repositories still needs to improve as they are below the average. Most of IRs in Zambia are not interoperable and cannot support federated services like local and global portals metadata harvesting. The development of the ETD portal is not unique to Zambia, as other countries like South Africa developed their ETD portal that provides access to a country-specific collection of ETDs with the primary responsibility of coordinating, managing, monitoring, and supporting the development of ETD programs at the various universities (Webley, Chipeperekwa & Suleman, 2011). The Zambian ETD portal for it to be successfully operational would require concerted efforts from HEIs by ensuring that IRs are configured in a way that they are interoperable. The need for interoperability among the IRs in the Global south was equally noted by Phiri (2018) as he emphasised the need to have IRs that could have specialised metadata schema with enough and correct metadata elements that would enable automatic harvesting of metadata from local to global portals for improved visibility of ETDs.

All the HEIs that have implemented the IRs in the management and organisation of the ETDs use general metadata schema, i.e., DSpace software with Dublin Core metadata schema, despite having ETD-MS, the schema designed specifically for ETDs. The bias to the Dublin Core metadata schema was due to its integration in DSpace, causing the librarians to adopt it as a default scheme. At the same time, the Librarians lacked knowledge about the existence of a specialised Scheme for ETDs (ETD-MS). The implication of non-compliance and having incomplete metadata elements is that most elements cannot be mapped correctly to ETD-MS schema while downstream services automatically harvest metadata; this limits the visibility of ETDs in Zambia.

Research also observed that most functional IRs from HEIs did not configure OAI-PMH. This hinders the interoperability among repositories. The implication is that most repositories cannot be registered by major databases like NDLTD Union Catalogue, which harvests ETD metadata from all registered open-access repositories. However, the variations across all IRs does not affect the local IR, the issue arise only within a collaborative network. As IRs move towards collaboration, it is vital that they all adhere to same standards as variations affect the interoperability in federated environment. This negatively affects the visibility of the ETDs, consequently affecting HEIs' ranking profiles at the national, continental, and world levels. The findings agree with earlier studies by Bong and Ebrahim (2017), who observed that numerous universities ranking systems directly or indirectly relying on research visibility and impact.

5.3. Quality of Metadata in Higher Education Institutions in Zambia

The quality of metadata, as presented in section 4.3 of chapter 4, indicates that the quality of ETD metadata in HEIs could have been more comprehensive as they lacked some critical elements compared to the de facto metadata schema for ETDs. In most repositories, batched elements like dc.contributor.author present a severe interoperability metadata standard issue, as the element is neither a recognised Dublin Core element nor an ETD-MS element. The use of dc.contributor.author is against

the International Standard Organization (ISO 15836-2:2019(en), a reference for the Dublin Core metadata element for description. This guides and describes all Dublin Core metadata elements to describe resources.[18]

The study also revealed that all the repositories in HEIs had missing metadata elements. The missing values and elements in all IRs in Zambia indicated that all institutions to some extent compromised the metadata ingestion process and quality control measures. The compromise was mostly due to lacking of knowledge about the de facto metadata scheme for ETDs amongst the policy makers in Libraries. The above findings are similar to Park and Richard (2011) in their study on Canadian IRs where they observed that 5 ETD-Ms elements were not being used in all 10 IRs studied. The research also revealed that null values in some repositories were partly due to a need for more workforces in all IRs to conduct the ingestion process as it is a labour-intensive undertaking. For example, despite the significant number of materials required for ingestion in the repository, only two Librarians were responsible for ingesting ETD metadata in UNZA IR.

Despite variations on missing metadata elements, the research also identified dc.contributor and theses.degree as two metadata elements constantly missing in all IRs. Much as it could be understood that not all elements could be used in all IRs, it concerns that dc.contributor and theses.degree are not being used at all in all IRs. The status quo demands answers as to whether elements included by ETD-Ms are extraneous or not. A summary metadata element used are shown in Appendix E.

5.4. Sources of Missing ETD-MS Metadata from Higher Education Institutions Repository in Zambia

This was made possible by identifying the sources of all metadata elements from ETDs by interviewing the research coordinators from all HEIs, and analysing regulations for postgraduate trainings handbooks together with ETD manuscripts.

---

[18] https://www.iso.org/standard/71341.html

The results from section 4.4 indicated that UNZA, MU and UNILUS have postgraduate handbooks that guides students on formatting the ETDs across all schools. These documents ensures that, students' manuscripts conform to the agreed structure and format before they are published. While an investigation into ZCAS, CHAU and LAMU showed that they do not have an explicit knowledge which outlines the required format for preparations of ETDs. The absence of policy documents that regulates postgraduate training mostly results in disorderly ways of preparing ETDs and eventually it becomes difficult to conduct proper examination of ETDs prior to publishing. At the same time, this results into interoperability challenges for downstream services both at National and at international level. The lack of guidelines affects the development of effective procedures for devising a natural language processing techniques that harvest missing ETD-MS metadata elements for improved visibility and interoperability between systems. This is in agreement with Ferreras et al. (2013) who observed that in order to promote cross-institutional collaboration and increase the visibility of Open Education Research of deposited educational resources be assigned with adequate metadata that assures effective retrieval and interoperability of IRs.

After undertaking the document analysis of policy documents and ETD manuscripts, the significant outcomes of possible elements drawn from the analysis indicated that the metadata elements were located on diverse sections of the preliminary pages of the manuscripts across IRs, except for CHAU and UNZA who have same standard outline for ETDs. Despite having variations on location of ETD-MS elements across HEIs, the results showed that there is uniformity within HEI in the way ETDs are prepared. Hence, developing an intelligent system using natural language processing to automatically generate the missing ETD metadata elements in Zambia would be possible and an outstanding achievement towards advancing the metadata quality for improved visibility of ETDs and interoperability of the IRs in Zambia and this would ensure a success to the launched Zambian ETDs portal. In a similar case, Phiri (2021) proposed to use the "text on the title page as text features that would help determine the type of ETD." This was after analysing the ETDs' preliminary pages from UNZA

IR to automatically classify digital objects for improved metadata quality. He observed that the location of the title information was on the upper text on the title pages. This is similar to Nyambe et al (2022); Azimjonov and Alikhanov (2018) who engineered a rule-based system for metadata extraction from Portable Document Format (PDF). The researchers applied general layout rules, such as "title element is usually located on the upper portions of the first pages, and they are usually in the largest font sizes," this represented a common approach aimed at recognizing and extracting specific text from a chosen format. Ongoing metadata extraction research often targets specific documents within the same file format. The variations that exist in location of missing ETD-Ms metadata entails sorting out incomplete and poor-quality metadata is possible, but challenging as it requires designing different sets of systems that could automatically generate missing ETD-MS metadata for every IR.

5.4.1. Summary of Chapter 5

The chapter has emerged that the implementation IRs in Zambia is very low and still need to be enhanced. These findings were consisted with other scholars who in their research found that implementation of IRs in Africa as a whole was still in infancy stage.

The performance of the implemented IRs is below the average. And majority of them are not interoperable and cannot support federated services like local and global portals metadata harvesting. The launch and implementation of ZANETD portal is not new as other countries like South Africa developed their ETD portal that provides access to a country-specific collection of ETDs at the national level.

There is no single IR in Zambia that uses ETD-MS metadata scheme despite being the specialised schema that was designed specifically for ETDs. The bias to the Dublin Core metadata schema according to research was due to its integration in DSpace and partly due to lack of lack of knowledge about the existence of a specialised Scheme for ETDs amongst information professionals in Zambia.

The implication of non-compliance is that it hampers interoperability as most metadata elements cannot be mapped correctly to ETD-MS schema while downstream services automatically harvest metadata. This limits the visibility of ETDs in Zambia as most repositories cannot be registered by major databases like NDLTD Union catalogue, which harvests ETD metadata from all registered open-access repositories and consequently affecting HEIs' ranking profiles.

The lack of uniformity with regards to standard outline and format of ETDs characterise most IRs in Zambia. The results shown variations in locations of ETD-MS metadata elements used in various HEIs. However, application of NLP for automatically extraction of missing ETD-MS metadata elements possible. The Appendix F shows variations metadata of the missing ETD-MS metadata elements.

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

6.1. Overview

The research pinpointed out the HEIs in Zambia that have operational and interoperable IRs. It sought to evaluate the comparative quality of ETD metadata within HEIs and identify potential sources of missed ETD metadata originating from HEIs. Consequently, based on this context, the chapter was poised to draw conclusions and provide recommendations arising from the study.

6.2. Conclusion

The increased rate at which digital objects are generated demands best practices in metadata generation that would see improved visibility and access to digital objects. Digitization of information resources does not result in improved visibility and access to digital objects when metadata generation is compromised. Therefore, IR managers must put in measures that ensures the generation of high-quality metadata as they have broad implications for downstream services that automatically harvest ETDs for dissemination purpose.

6.2.1 Higher Education Institution that Have Functional and Interoperable Online Institution Repositories in Zambia

The research findings have shown that most HEIs are yet to implement IRs in Zambia as only 10 out of 61 HEIs have adopted and implemented IRs in Zambia. The implemented IRs needs to be reconfigured by activating OAI PMH as most of them are not interoperable and cannot support federated metadata harvesting by local and global ETD portals as it was discovered that only 2 out of the 10 IRs activated the OAI PMH.

The low the number of HEIs that implemented the IRs in Zambia, the low the visibility of research output, which eventually affects the ranking of the HEIs locally and internationally and to some extent, it negatively affects the fight against academic dishonesty or plagiarism.

Further, it has been established that all the 10 HEIs that implemented the IRs in the management and organisation of the ETDs use DSpace with Dublin Core metadata schema, despite having ETD-MS schema designed specifically for ETDs.

6.2.2 Quality of Metadata in Higher Education Institutions in Zambia

Research found that all IRs were not compliant to ETD-MS, the defacto metadata schema developed by NDLTD.

Further, it was established that metadata ingestion process and adopted scheme plays a vital role in maintaining high quality for all digital objects. The metadata ingestion process being a labour-intensive task, requires balanced workforce against workload. The compromised metadata quality control at all level is attributed to the presence of null values in some used elements in IRs.

6.2.3 Potential Sources of Missing Electronic Theses Dissertations Metadata Originating from Higher Education Institutions in Zambia

Research revealed that the location of missing ETD MS varied from each HEI. For instance, dc.contributor details were found on the title page of most manuscripts for most HEIs, these details were not available in manuscripts for some HEIs. However, all IRs maintained homogeneity with regards to ETD format within an institution. The identification of sources of missing ETD metadata, presents an opportunity for automatically generation of missing ETD MS metadata using Natural Language Processing.

6.3. Recommendations

Having understood the strength and weaknesses of IRs in Zambia, some of the recommendations have been proposed which include:

i. All IRs must conform to the same metadata standard that enables metadata harvesting by national and global ETD portal.

ii.   There is need for IRs to be compliant to the de facto metadata schema, the ETD-MS for local and international collaborations

iii. There is need to design the intelligent system that would automatically generate the missing ETD metadata elements for improved visibility of research output

REFERENCES

Adam, U. (2021). Institutional repositories in Africa: regaining direction. *Journal of Information Science, 47*(6), 780-795. https://doi.org/10.1177/02666669211015429

Alemneh, D. (2009). Metadata Quality: A phased approach to ensuring long-term access to digital resources. *Proceedings of the American Society for Information Science and Technology, 46(1),* 1-6. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc29318/.

Azimjonov, J., & Alikhanov, J. (2018). Rule-based metadata extraction framework from academic articles. ArXiv, 1807-9009. Retrieved from https://api.semanticscholar.org/CorpusID:50777379

Baker, S. E., & Edwards, R. (2012). How many qualitative interviews is enough?: Expertvoices and early career reflections on sampling and cases in qualitative research. National Centre for Research Methods Review Paper. Retrieved from http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf

Betz, S., & Hall, R. (2015). Self-archiving with ease in an institutional repository: Microinteractions and the user experience. *Information Technology and Libraries, 34*(3), 43-58.

Bong, Y. B., & Ale, E. N. (2017). Increasing visibility and enhancing impact of research. *Asia Research News.* Retrieved from http://www.researchsea.com/html/article.php/aid/10634/cid/1. Accessed on 27 December, 2022.

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. ALA Editions. Retrieved from *https://www.researchgate.net/publication/247818823_The_Continuum_of_ Metadata_Quality_Defining_Expressing_Exploiting*

Callicott, B., Scherer, D., & Wesolek, A. (2016). *Making Institutional Repositories Work.* Purdue University Press. https://doi.org/10.2307/j.ctt1wf4drg.4

Chan, C.H. (2008). Metadata Quality for Digital Libraries (Master's thesis). The Universityof Waikato, Hamilton, New Zealand. Retrieved from *https://hdl.handle.net/10289/2312*.

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

Coenen, M., Stamm, T. A., Stucki, G., & Cieza, A. (2012). Individual interviews and focus groups in patients with rheumatoid arthritis: A comparison of two qualitative methods. *Quality of Life Research, 21*(2), 359-370. doi:10.1007/s11136-011-9943-2.

Cohen, A. I. (2016). Ethics and professional practice. In E. Searing & D. Searing (Eds.), *Practicing Professional Ethics in Economics and Public Policy* (pp. 14). Springer, Dordrecht. *https://doi.org/10.1007/978-94-017-7306-5_14*

Creswell, J.W. (2013). Dual inquiry and research design: Choosing among five approaches. London: Sage.

Franzl, G., Wanzenböck, C., & Berger, A. (2023). Interoperability as a key factor for digitalisation — a success story for cross-sector knowledge transfer. *Elektrotech. Inftech., 140,* 471–477. https://doi.org/10.1007/s00502-023-01143-3

Soares, F., Maculan, B., & Drucker, D. (2019). Darwin Core for Agricultural Biodiversity: A metadata extension proposal. *Biodiversity Information Science and Standards, 3*. https://doi.org/10.3897/biss.3.37053

Enago Academy. (2021). Thesis vs. Dissertation. Retrieved on 2021-11-11 from *https://www.enago.com/academy/thesis-vs-dissertation/.*

Foster, N. F., & Gibbons, S. (2005). Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine, 11*(1). Retrieved from http://www.dlib.org/dlib/january05/foster/01foster.html

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health, 25(*10), 1229–1245. https://doi.org/10.1080/08870440903194015

Global Biodiversity Information Facility. (2019). What is Darwin Core, and why does it matter? Retrieved from https://www.gbif.org/darwin-core. Accessed on: 2022-04-24.

Gumpili, S. P., & Das, A. V. (2022). Sample size and its evolution in research. *IHOPE Journal of Ophthalmology, 1(*1), 9-13. doi:10.25259/IHOPEJO_3_2021

Hennink, M. A., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research:A systematic review of empirical tests. *Social Science & Medicine*, 292(8), 11452. https://doi.org/10.1016/j.socscimed.2021.11452. (Accessed on 15th March, 2022).

Higher Education Authority. (2019*). The state of higher education in Zambia: 5 decades of university education.* Lusaka: Higher Education Authority. Retrieved from https://hea.org.zm/sdm_downloads/the-state-of-higher-education-in-zambia-2019-5-decades-of-university-education/

Huber, S., Wiemer, J., Schneider, D., Ihlenfeldt, S. (2019). Data mining methodology for engineering applications a holistic extension to the CRISP-DM model, Procedia CIRP,79(4), 403-408, https://doi.org/10.1016/j.procir.2019.02.106. (Accessed on 24/12/2022).

Kothari, R., & Garg, G. (2014). *Research Methodology: Methods and Techniques* (3rd ed.). New Delhi: New Age International (P) Limited Publishers.

MacDonald, R. (2011). Building a digital repository program with limited resources. *The Electronic Library, 29*(6), 854-855. https://doi.org/10.1108/02640471111188097. (Accessed on 4/3/2022).

Milena, D., Yunhyong, K., & Seamus, R. (2013). Digital Curation Reference Manual Installment on "Automated Metadata Generation". Retrieved from http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction. (Accessed on: 14 April 2021).

Ministry of Higher Education. (2023). *Higher Education Amendment Act No. 23*. Government Printers.

Mishra, C., & Gupta, D. (2017). Deep Machine Learning and Neural Networks: An overview. *IAES International Journal of Artificial Intelligence (IJ-AI), 6*, 66. doi:10.11591/ijai.v6.i2.pp66-73.

Morse, W. C., Damon, R. L., & Todd, S. (2014). Exploring saturation of themes and Spatial locations in qualitative public participation geographic information systems research. *Society & Natural Resources, 27(*5), 557-571. https://doi.org/10.1080/08941920.2014.888791. (Accessed July 28, 2022).

National ETD Portal. (n.d.). Retrieved March 23, 2021, from http://lis.unza.zm/portal

NDLTD Union Archive. (n.d.). Retrieved March 23, 2021, from
http://union.ndltd.org

OATD – Open Access Theses and Dissertations. (n.d.). Retrieved March 23, 2021, from https://oatd.org

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital . *International Journal on Digital Libraries.* https://doi.org/10.1007/s00799-009- 0054-4. (Accessed on: 30 December 2021).

Ochoa, X. (2014). Metadata quality. In E. R. H. H. Loup (Ed.), *Handbook of Metadata, Semantics and Ontologies* (pp. 63–88).

Pal, J. K. (2010). Metadata initiatives and emerging technologies to improve resource discovery. *Annals of Library and Information Studies, 57*(12), 44-53.

Park, E. G., & Richard, M. (2011). Metadata assessment in e-theses and dissertations Of Canadian institutional repositories. *The Electronic Library, 29*(3), 394-407. (Accessed on 14 November 2023)

Park, J., & Tosaka, Y. (2010). Metadata quality control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly, 48*(8), 696-715. doi:10.1080/01639374.2010.508711

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. Available at: https://doi.org/10.1080/01639370902737240. (Accessed 03 August, 2021).

Park, W., & Song, J. (2019). Between realism and constructivism. In *The Routledge Companion to Qualitative Political Research* (pp. 213-226). Routledge. https://doi.org/10.4324/9781351214827-14

Phiri, L. (2018). Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia. In *Proceedings of the 2nd IEEE International Conference in Information and Communication Technologies (ICICT 2018),* Lusaka, Zambia. Retrieved from http://dspace.unza.zm/handle/123456789/5723. (Accessed 14 June 2021

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.

Potvin, S., & Thompson, S. (2016). An analysis of evolving metadata influences, standards, and practices in Electronic Theses and Dissertations. *Library Resources & Technical Services, 60(2),* 76-78. Retrieved from https://journals.ala.org/index.php/lrts/issue/view/544. (Accessed on 24/11/2022).

Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For? Retrieved from https://www.niso.org/publications/understanding-metadata-2017. (Accessed on: 25 August 2020).

Suleman, H. (2012). The NDLTD Union Catalog: Issues at a Global Scale. Retrieved from http://hdl.handle.net/10757/622568.

Suleman, H., & Fox, E. A. (2012). Leveraging OAI harvesting to disseminate theses. *Library Hi Tech, 21*(2), 219-227 https://doi.org/10.1108/07378830310479857. (Accessed on 23 August 2021).

Tahmed, M., Mukhopadhyay, M., & Mukhopadhyay, P. (2023). Automated Knowledge organization: AI/ML-based subject indexing system for libraries. *Journal of Library & Information Technology, 43*(1), 45-54. https://doi.org/10.14429/djlit.43.01.18619.

Tani, A., Candela, L., & Castelli, D. (2013). Dealing With Metadata Quality: The Legacy of Digital Library Efforts. *Information Processing & Management, 49*(6), 1194–1205.

The Networked Digital Library of Theses and Dissertations. (2015). ETD-MS v1.1: An Interoperability Metadata Standard for Electronic Theses and Dissertations. Retrieved from http://www.ndltd.org/standards/metadata. (Accessed 12 July, 2021).

UNESCO. (2014). Global Open Access Portal. Retrieved from https://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/access/-by-region/latin-america-and-the-caribbean/brazil/

Varshil Bhagaji Dashrath. (2014). Role of Metadata in Digital Resource Management. *International Journal of Digital Library Services (IJODLS),* 4(3), 209-214. Retrieved from http://www.ijodls.in/uploads/3/6/0/3/3603729/19434.pdf

Vasileiou, K., Barnett, J., & Thorpe, S. (2018). Characterizing and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology, 18(I4), 148. https://doi.org/10.1186/s12874-018-0594-7.*

Velmurugan, C. (2010). Institutional Repositories: A powerful tool for accessing information for Educationalists. In *Proceedings of the National Conference on* Next Generation Digital Libraries and Web Technologies: Challenges and Opportunities (pp. 6.92-6.97). Sri Krishna College of Engineering & Technology, Coimbatore**.**

Velmurugan, C. (2013). Open-source software: An institutional digital repository system with special reference to DSpace software in digital libraries - an introduction. *International Journal of Library and Information Science, 5*(10), 313-318. Retrieved from http://www.academicjournals.org/IJLIS.

Weagley, J., Gelches, E., & Park, R. (2010). Interoperability and Metadata Quality in Digital Video Repositories: A Study of Dublin Core. *Journal of Library Metadata, 10*(1), 15-43.

Wieczorek, J., Bloom, D., Guralnick, R., & Stan, B. (2012). Darwin Core: An

evolving Community-Developed Biodiversity Data Standard. *PLoS ONE, 7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715. (Accessed on 23 November 2021).

Yusuf, F., Ifijeh, G., & Emmanuel, O. (2019). Institutional repositories in Africa: Issues and challenges. In *Handbook of Research on Emerging Trends and Technologies in Library and Information Science* (pp. 123-140). doi:10.4018/978-1-5225-8437-7.ch008.

Appendices

Appendix A. Article: Towards metadata completeness in national ETD portal for
improved discoverability.

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.

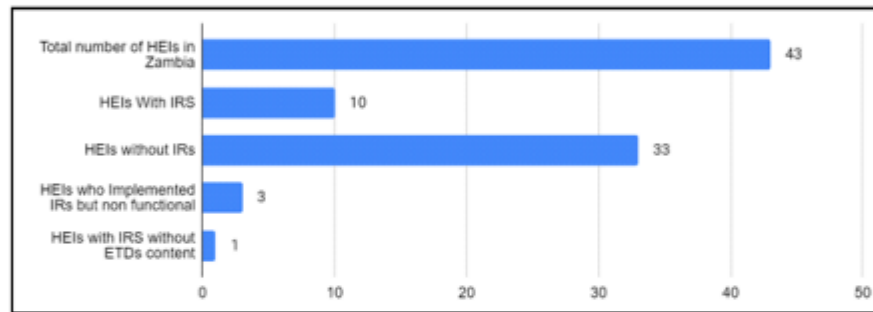**Figure 1: Status of institutional Repositories in Zambia**

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.
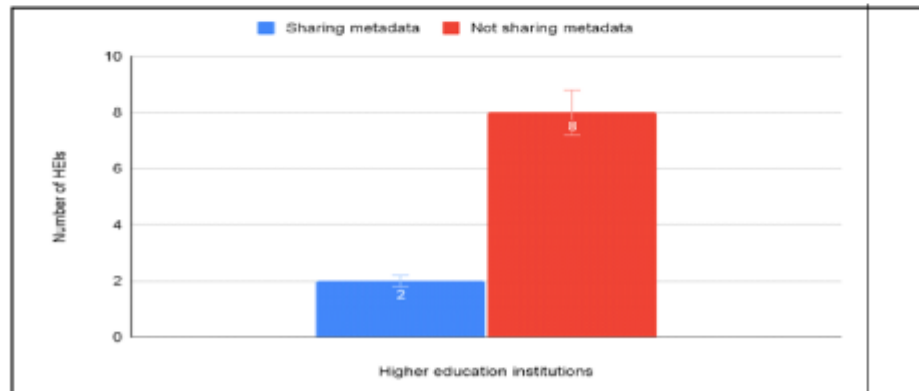

Figure 2: Interoperability of Institutional Repositories in Zambia

- 523 -

94

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.

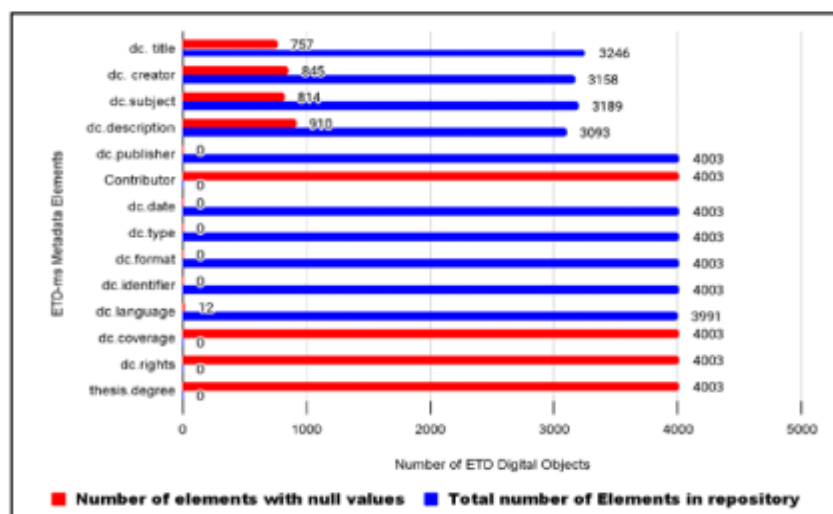**Figure 3: Metadata elements used at UNZA IR**

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.

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.
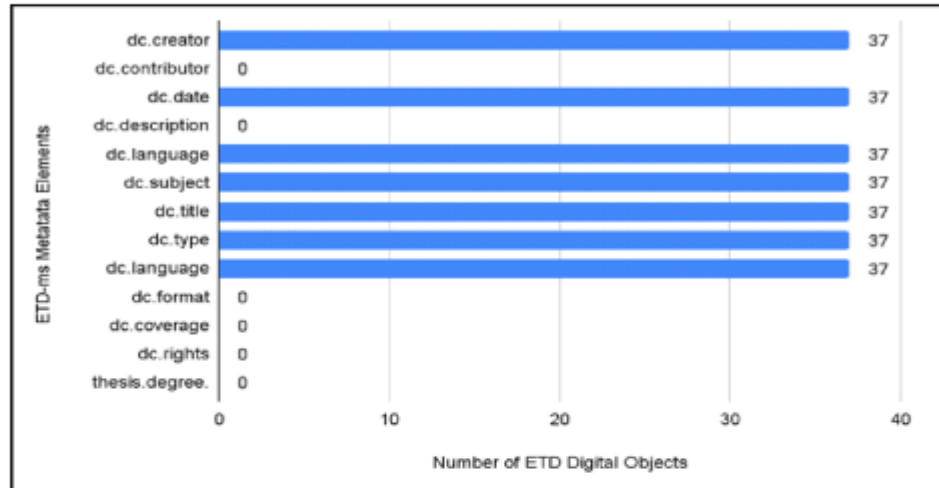

Figure 4: Metadata elements used at UNILUS Vs ETD-MS

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/

- 526 -

98

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.

Empirical Evaluation of ETD-ms Compliance in Institutional Repositories

There are presently a number of academic institutions that have set up Institutional Repositories (IRs) aimed, in part, to archive, manage and facilitate access to Electronic Theses and Dissertations (ETDs). Increasingly, descriptive metadata associated with the ETDs produced by most of these institutions are now automatically harvested by National ETD portals and entities such as The Networked Digital Library of Theses and Dissertation (NDLTD) Union Catalog. However, there are growing concerns regarding the quality of ETD descriptive metadata originating from most of these institutions, in particular, adherence to the ETD-ms metadata standard, the de-facto standard used to describe ETDs. Compliance with well-established metadata standards such as ETD-ms (Hickey et al., 2021) can arguably improve the quality of metadata.

Suleman highlights that errors in metadata records harvested by the NDLTD Union Catalog are a recurring problem and that while widely advertised, ETD-md compliance is a major challenge (Suleman, 2012).

The quality of metadata is known and cited as being crucial to the effective preservation of digital content and, more significantly, aids in the discoverability of digital content (Park, 2009). There are a number of studies that have been conducted in order to investigate metadata quality. Studies such as the one conducted by Currier et al. have been used to identify errors in learning object repositories, object metadata, by untrained resource creators and the lack of use of authority control and subjects as being the major issues associated with quality of metadata (Currier et al., 2004). In addition, bibliometric analyses of the Dryad repository have identified major problems with Creator, Data and Type metadata elements (Rousidis et al., 2014).

This paper presents an empirical analysis of ETD-ms compliance of the metadata associated with the 5,949,744 ETDs in the NDLTD Union Catalog. ETD metadata records were harvested from the NDLTD Catalog using the OAI-PMH protocol and, subsequently, an analysis was conducted to determine compliance to the ETD-ms metadata standard. In addition, to provide contextual overview of the potential root causes of non-compliance to ETD-ms, a case study was conducted at a Higher Education Institution, in order to understand current practices and procedures employed during ingestion and association of metadata quality. The study could potentially provide direction on issues to be addressed during the ingestion of ETDs into IRs

References

Currier, S., Barton, J., O'Beirne, R., & Ryan, B. (2004). Quality assurance for digital learning object repositories: issues for the metadata creation process. In *Research in Learning Technology* (Vol. 12, Issue 1). https://doi.org/10.3402/rlt.v12i1.11223

Hickey, T., Pavani, A., & Suleman, H. (2021, April). *ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations.* https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html

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

Rousidis, D., Garoufallou, E., Balatsoukas, P., & Sicilia, M.-A. (2014). Data Quality Issues and Content Analysis for Research Data Repositories: The Case of Dryad. In *Let's Put Data to Use: Digital Scholarship for the Next Generation* (pp. 49–58). IOS Press.

Suleman, H. (2012). *The NDLTD Union Catalog: Issues at a Global Scale.* http://hdl.handle.net/10757/622568

Appendix C. Survey Questionnaire

Dear Respondent,

I am a Master of Library and Information Science student at the University of Zambia, Great East Road campus researching on Automatic Generation of Electronic Theses and Dissertations Metadata.

Please be assured that the information you will provide in this survey is purely for academic purposes only and will be treated with the utmost confidentiality. Your participation will be highly appreciated.

Yours sincerely,

Adrian Chisale

Contact:

 0979228493

chisalea1986@gmail.com

(1). Select the institution you belong to below.

…………………………………………………………………………………… ……………

(2). Do you offer post-graduate study programs at your Institution?

1. YES

2. NO

(3). If your response to question number (2) above is YES, how do you manage Thesis and dissertation in your organization?

………………………………………………………………………………… ……………....

(4). Do you have an online institutional repository in your organization?

a. NO

b. YES

If your response to question number (4) is NO, skip all the remaining questions by scrolling down and submit your questionnaire when you are done.

(5). Do you manage electronic theses and dissertations in your repository?

a. yes

b. No

(6). On what platform (software) is your repository installed? (e.g DSppace, E-prints etc)

………………………………………………………………………………………

……..

(7). What is the web address/ Uniform Resource Locator for your Institutional Repository? (e.g http://dspace.unza.zm/)

………………………………………………………………………………..

(8). Is your platform allowing sharing of metadata among repositories?

a. yes

b. No

Other:

(9). What metadata schema is used to describe digital objects in your repository (e.g MARC, DC etc)?

………………………………………………………………………………………

…………….

(10). What protocols have you activated that allows sharing of metadata among repositories?

………………………………………………………………………………………

………………

THE END

Appendix D: A Live webinar on making Electronic Theses and Dissertations Portal in Zambia a Success

Appendix E: Interview Guide

1. Kindly state your role as research coordinator.
2. Do you have an explicit document guiding how ETDs must be presented with regards to format and outline? (YES / NO).
3. If the answer to question 2 is NO. How do you maintain ETDs' standards with regard to format and outline

Appendix F: University of Zambia Humanities and Social Sciences Research Ethics Committee Ethical Clearance Approval

THE UNIVERSITY OF ZAMBIA
DIRECTORATE OF RESEARCH AND GRADUATE STUDIES

Great East Road Campus | P.O. Box 32379 | Lusaka10101 | Tel: +260-211-290 258/291 777 **Fax**: (+260)-211-290 258/253 952 | E-mail: director.drgs@unza.zm | Website: www.unza.zm

APPROVAL OF STUDY

IORG No. 0005376
HSSREC IRB No. 00006464

30th March, 2022

REF NO. HSSREC-2022-MAR-037

Adrian Chisale
The University of Zambia
IDE
P.O. Box 32379
LUSAKA

Dear Mr. Chisale,

RE: **"AUTOMATIC GENERATION OF ELECTRONIC THESES AND DISSERTATIONS METADATA"**

Reference is made to your submission of the protocol captioned above. The HSSREC resolved to approve this study and your participation as Principal Investigator for a period of one year.

REVIEW TYPE	ORDINARY REVIEW	APPROVAL NO. HSSREC-2022-MAR-037
Approval and Expiry Date	Approval Date: 30th March, 2022	Expiry Date: 29th March, 2023
Protocol Version and Date	Version - Nil.	29th March, 2023
Information Sheet, Consent Forms and Dates	☐ English.	To be provided
Consent form ID and Date	Version - Nil	To be provided
Recruitment Materials	Nil	Nil
Other Study Documents	Questionnaire.	
Number of Participants Approved for Study		

Specific conditions will apply to this approval. As Principal Investigator it is your responsibility to ensure that the contents of this letter are adhered to. If these are not adhered to, the approval may be suspended. Should the study be suspended, study sponsors and other regulatory authorities will be informed.

Conditions of Approval

- No participant may be involved in any study procedure prior to the study approval or after the expiration date.

- All unanticipated or Serious Adverse Events (SAEs) must be reported to HSSREC within 5 days.

- All protocol modifications must be approved by HSSREC prior to implementation unless they are intended to reduce risk (but must still be reported for approval). Modifications will include any change of investigator/s or site address.

- All protocol deviations must be reported to HSSREC within 5 working days.

- All recruitment materials must be approved by HSSREC prior to being used.

- Principal investigators are responsible for initiating Continuing Review proceedings. HSSREC will only approve a study for a period of 12 months.

- It is the responsibility of the PI to renew his/her ethics approval through a renewal application to HSSREC.

- Where the PI desires to extend the study after expiry of the study period, documents for study extension must be received by HSSREC at least 30 days before the expiry date. This is for the purpose of facilitating the review process. Documents received within 30 days after expiry will be labelled "late submissions" and will incur a penalty fee of K500.00. No study shall be renewed whose documents are submitted for renewal 30 days after expiry of the certificate.

- Every 6 (six) months a progress report form supplied by The University of Zambia Humanities and Social Sciences Research Ethics Committee as an IRB must be filled in and submitted to us. There is a penalty of K500.00 for failure to submit the report.

- When closing a project, the PI is responsible for notifying, in writing or using the Research Ethics and Management Online (REMO), both HSSREC and the National Health Research Authority (NHRA) when ethics certification is no longer required for a project.
- In order to close an approved study, a Closing Report must be submitted in writing or through the REMO system. A Closing Report should be filed when data collection has ended and the study team will no longer be using human participants or animals or secondary data or have any direct or indirect contact with the research participants or animals for the study.

- Filing a closing report (rather than just letting your approval lapse) is important as it assists HSSREC in efficiently tracking and reporting on projects. Note that some funding agencies and sponsors require a notice of closure from the IRB which had approved the study and can only be generated after the Closing Report has been filed.

- A reprint of this letter shall be done at a fee.

- All protocol modifications must be approved by HSSREC by way of an application for an amendment prior to implementation unless they are intended to reduce risk (but must still be reported for approval). Modifications will include any change of investigator/s or site address or methodology and methods. Many modifications entail minimal risk adjustments to a protocol and/or consent form and can be made on an Expedited basis (via the IRB Chair). Some examples are: format changes, correcting spelling errors, adding key personnel, minor changes to questionnaires, recruiting and changes, and so forth. Other, more substantive changes, especially those that may alter the risk-benefit ratio, may require Full Board review. In all cases, except where noted above regarding subject safety, any changes to any protocol document or procedure must first be approved by HSSREC before they can be implemented.

Should you have any questions regarding anything indicated in this letter, please do not hesitate to get in touch with us at the above indicated address.

On behalf of HSSREC, we would like to wish you all the success as you carry out your study.

Yours faithfully,

Dr. J. I. Ziwa

DR. J. I. Ziwa

ACTING CHAIRPERSON
THE UNIVERSITY OF ZAMBIA HUMANITIES AND
SOCIAL SCIENCES RESEARCH ETHICS COMMITTEE - IRB

cc: Director, Directorate of Research and Graduate Studies
     Assistant Director (Research), Directorate of Research and Graduate Studies
     Assistant Registrar (Research), Directorate of Research and Graduate Studies

Appendix G: Location of Dc.Contributor element From ZCAS University.

ZCAS UNIVERSITY

UNIVERSITY
of
GREENWICH

ZCAS
UNIVERSITY

BUSINESS SCHOOL, AUGUST 2016 STARTER

PROGRAMME:	MASTERS OF BUSINESS ADMINISTRATION
STUDENT NUMBER:	000589192
COURSE:	BUSI 1359 – PATHWAY SPECIALISATION PROJECT
TITLE:	ENTREPRENUERSHIP IN ZAMBIA: AN INVESTIGATION OF THE LOW SURVIVAL RATE OF MSMEs AND ITS EFFECT ON ZAMBIA'S ECONOMY
LECTURER:	DR. KWESI SAKYI
DUE DATE:	26.09.2018
WORD COUNT:	10 150

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT FOR THE AWARD OF MBA IN INTERNATIONAL BUSINESS AT THE UNIVERSITY OF GREENWICH BUSINESS SCHOOL.

ACKNOWLDGEMENTS

I would first like to appreciate and thank my thesis supervisor Dr. Kwesi Sakyi, senior lecturer in the School of Business Faculty at Zambia Centre for Accountancy Studies University (ZCAS) for his sacrifice and guidance during the entire period of me doing this thesis. I would also like to thank Dr. Richard Mbewe from whom I have obtained a wealth of knowledge throughout the duration of my course.

I am also grateful to the small business owners who took time to respond to my survey as well as speak to me concerning their business operations. Finally I would like to thank my young brother Wankunda Simuchimba for helping me during my data collection.

Above all I thank Almighty God from who all good things come from for giving me this opportunity to do this thesis and for sustaining me both physically and spiritually as I undertook this mammoth task.

Appendix H: Location of dc.contributor Element From the University of Zambia

PERCEPTIONS OF NURSES AT NDOLA TEACHING HOSPITAL TOWARDS SEXUAL HEALTH NEEDS OF PEOPLE WITH MENTAL HEALTH PROBLEMS

BY

CHAMBATU ALICE (RN, RNM, BSc)

A dissertation submitted in partial fulfilment of the requirements for the Degree of Master of Science in Midwifery and Women's Health

THE UNIVERSITY OF ZAMBIA

LUSAKA

March, 2022

ACKNOWLEDGEMENTS

I wish to acknowledge the lecturers at the University of Zambia, School of Nursing Sciences for equipping me with knowledge in the field of my study as well as with knowledge on how to conduct research.

Particularly, I acknowledge my research supervisors; Dr C.N. Kwaleyela (Main supervisor) and Mrs K.C. Mutambo (Co-supervisor) for guiding me in the writing of this research.

Lastly I acknowledge my family and friends for the encouragement and help rendered to me during the process of the research proposal development.

Appendix I: Location of Dc.contributor Element from Chalimbana University.

CHALIMBANA UNIVERSITY
DEPARTMENT OF DISTANCE EDUCATION
EDUCATION LEADERSHIP AND MANAGEMENT

COMPUTER NO.:5201920099

COURSSE CODE: MELM

YEAR OF STUDY: SECOND COHORT

ADMINISTRATIVE EFFECTS OF STEREOTYPES AND LABELLING ON FEMALE PUPIL'S ACADEMIC ACHIEVEMENT AT GRADE 12. A CASE STUDY OF SECONDARY SCHOOLS IN NAKONDE DISTRICT, ZAMBIA.

BY

SYLVIA NKOLE NSELUKA

A DISSERTATION SUBMITTED TO CHALIMBANA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF EDUCATION IN EDUCATION LEADERSHIP AND MANAGEMENT

2

Appendix Differences and Similarities of all the Metadata Elements used in Institutional
Repositories in Zambia

Similarities and differences of metadata elements used in Higher education institutions in Zambia							
metadata elements	CAU	MU	ZCAS	CHAU	LAMU	UNZA	UNILUS
dc.creator	✔	✔	✔	✔	✔	✔	✔
dc.contributor	X	X	X	X	X	X	X
dc.date	✔	✔	✔	✔	✔	✔	✔
dc.format	X	✔	X	X	X	X	X
dc.identifier	✔	✔	✔	✔	✔	✔	✔
dc.description.	✔	✔	✔	✔	✔	✔	X
dc.coverage	X	✔	X	X	X	X	X
dc.language	✔	✔	✔	✔	X	✔	X
dc.publisher	✔	✔	✔	X	✔	✔	X
dc.subject	✔	✔	✔	✔	✔	✔	✔
dc.title	✔	✔	✔	✔	✔	✔	✔
dc.type	✔	✔	✔	✔	✔	✔	✔
dc.rights	X	X	✔	X	X	X	X
thesis.degree	X	X	X	X	x	X	X

Appendix K. Variations on some ETD-MS Elements

Higher Education Institutions	ETD-MS Metadata element missing	Location of missing metadata in ETD (Sections of ETD)
University of Zambia IR	dc.contributor (Supervisor/ Advisor details)	Acknowledgements/ certificate of approval
Mulungushi University IR		Title page/ Supervisor's Recommendation/ Acknowledgement
Chalimbana University IR		Approval/ Acknowledgement/ Declaration
Cavendish University IR		Title page/ Declaration/ Acknowledgement/
Lusaka Apex Medical University IR		Title page/ Supervisor's declaration