The Zambia National Electronic Theses and Dissertations Pre-Processing Pipeline Portal

By

Chilufya Chanda (2019020963)

Chileshe Kamfwa (2019084244)

Chuulu Mainda (2019067218)

Supervisor:

Dr. Lighton Phiri

A report submitted to the Department of Library and Information Science, The University of Zambia, in partial fulfillment of the requirements of the degree of Bachelor of Information and Communication Technologies with Education

THE UNIVERSITY OF ZAMBIA

LUSAKA

2023

Abstract

Higher Education Institutions (HEIs) in Zambia contribute significantly to scholarly research through the production of theses and dissertation manuscripts. These scholarly works are archived in Institutional Repositories (IRs) and aggregated in the Zambia National Electronic Theses and Dissertations (ETD) portal. However, inconsistencies in metadata formatting and missing metadata fields hinder the seamless accessibility of these valuable resources. This thesis presents a novel approach to addressing these challenges by developing a pre-processing pipeline software tool that harnesses existing machine learning models. The tool aims to ensure the consistent formatting of metadata and enhance missing metadata, thereby improving the overall usability and accessibility of the ETD portal.

Acknowledgements

We would like to express our heartfelt gratitude to all those who have contributed to the successful completion of our final year project, "The Electronic Theses and Dissertations Preprocessing Pipeline Portal." This project would not have been possible without the support, guidance, and encouragement of many individuals and organizations.

First and foremost, we would like to thank our project supervisor who is also the ICT 4014 Coordinator, who has been an invaluable source of knowledge, wisdom, and guidance throughout this journey. Your mentorship and expertise in the field of Information and Communication Technology have been instrumental in shaping the direction of this project.

We extend our appreciation to the University of Zambia for providing us with the necessary resources and infrastructure to conduct this research. The access to the university's library, academic databases and network significantly contributed to the success of this project.

We would also like to acknowledge our fellow students who offered their support, insights, and feedback during the various stages of this project. Your collaborative spirit and shared enthusiasm for this field have been a source of motivation.

We cannot overlook the importance of the existing machine learning models that formed the backbone of this project. Their contribution in augmenting and formatting metadata consistently is noteworthy.

Lastly, we would like to express our gratitude to our friends and family for their unwavering support and understanding throughout this challenging yet rewarding journey.

In conclusion, this project has been a significant learning experience, and the knowledge and skills we have gained will undoubtedly benefit us in our future endeavors in the field of Information and Communication Technology.

Thank you to all for being a part of this incredible journey.

Project Team 14

Table of Contents

Table of Contents

CHAPTER SIX

List of Tables

List of Figures

List of Abbreviations

Abbreviations	**Description**
API	Application Programming Interface
ETD	Electronic Theses and Dissertations
HEIs	Higher Education Repositories
IRs	Institutional Repositories
NETD	National Electronic Theses and Dissertations
NDLTD	Networked Digital Library of Theses and Dissertations
OAI-PMH	Open Archives Initiative for Metadata Harvesting
MU	Mulungushi repository
UNZA	University of Zambia
ZOU	Zambia Open University

CHAPTER ONE

1. Introduction

1.2 Background

Zambia's HEIs contribute to the global academic landscape by producing a wealth of research in the form of theses and dissertations. The aggregation of these works in the Zambia National ETD portal has proven invaluable for researchers and scholars seeking comprehensive access. However, the presence of inconsistent metadata formatting and incomplete metadata fields poses obstacles to efficient resource discovery

1.3 Problem statement

Zambian HEIs offer advanced postgraduate programs that result in the publication of theses and dissertation manuscripts. The ETDs are archived and made available via Institutional Repositories (IRs). Downstream services such as National ETD portals and global portals are generally used to aggregate metadata originating from such IRs. Inconsistencies in metadata formatting and missing metadata are common challenges that prevent end-users from accessing ETDs effectively. There is, therefore, a need to build a pre-processing pipeline software tool that will consistently format metadata and augment missing metadata.

1.4 Objectives of study

1.4.1 Broad Objective

The broad objective of this project is to improve the accessibility of ETDs generated by Zambian HEIs by building a pre-processing pipeline software tool that will consistently format metadata and augment missing metadata using existing machine learning models.

1.4.2 Specific Objectives

1. To investigate the inconsistencies of metadata from institution repositories
2 .To design and implement the pre-processing pipeline software tool for ETD metadata
3. To demonstrate the completeness and consistency of Pre-processed ETD metadata .
4.To have a unified interface for the presentation of the ETDs

1.5 Research Questions

1. Is the software tool capable of efficiently handling diverse types of metadata?
2. Will the software tool format and ensure accurate extraction and transformation of information?
3. How will the effectiveness of the software tool be assessed?

1.6 Ethical Considerations

When discussing the pre-processing pipeline for Electronic Theses and Dissertations (ETDs), there are several ethical considerations to keep in mind. These considerations are centered on the handling and processing of sensitive or confidential data, ensuring data privacy and security, and adhering to ethical guidelines. Here are some key ethical considerations for an ETD pre-processing pipeline:.

1.6.1 Intellectual property

The project will respect the intellectual property rights of the owners of the ETDs. The harvesting and use of ETD metadata should be done in accordance with copyright laws and regulations.

1.6.2 Data Security

The software will ensure that the metadata collected and processed is secured against unauthorized access, modification, or disclosure. Appropriate security measures will be put in place to protect the integrity and confidentiality of the data.

CHAPTER TWO

2. Related Work

In the rapidly evolving landscape of scholarly communication, Electronic Theses and Dissertations (ETDs) play a pivotal role, offering a treasure of knowledge and insights from academic research. As these valuable resources continue to proliferate in the digital realm, the need for effective access, discovery, and dissemination becomes increasingly pronounced. To address these demands, various initiatives have emerged, aiming to centralize and facilitate the exploration of ETDs through dedicated portals. This chapter presents an exploration of two prominent ETD portal initiatives—the South African National Electronic Thesis[1] and Dissertation (NETD) portal and the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog[2]. Each initiative addresses unique challenges related to downstream services, portal architecture, metadata, and metadata quality.

2.1 Metadata and Metadata Quality

Metadata refers to the structured information about data, encompassing details such as its origin, format, structure, and content. It acts between users and data, enabling efficient discovery, understanding and utilization of information.[1]. Metadata can exist in various forms including bibliographic metadata for books, technical metadata for digital files, and so on.

Metadata Quality refers to the accuracy, completeness,consistency and relevance of metadata. High-quality metadata can lead to erroneous decisions and wasted resources. Ensuring metadata quality is crucial for organizations to maximize the value of their data assets.

To achieve metadata quality, organizations should adhere to metadata standards and best practices. The adoption of standards like Dublin core for general metadata ensures consistency and interoperability. Automation tools can assist in maintaining accuracy and consistency over time.

Both the NETD and NDLTD Union Catalog initiatives, which are the center of our related work, emphasize the critical role of metadata quality. Metadata serves as the bridge between users and ETD resources, facilitating effective discovery and access. The quality of metadata profoundly impacts discoverability, accessibility, and usability of ETDs. Ensuring accurate, standardized, and complete metadata becomes paramount for maintaining user trust and enhancing the scholarly impact of ETD repositories.

2.1.1 Dublin core elements

Our project utilizes the Dublin Core Metadata Initiative (DCMI) which outlines a series of straightforward and universal metadata elements designed for characterizing online resources. These elements serve as a foundation and framework for describing resources[7]. It initially consists of 15 elements, the Dublin Core Metadata Element Set has developed over time, maintaining its core components while incorporating extra elements and qualifiers and ensuring a versatile and effective foundation for describing a variety of web-based resources. Below are the 15 Dublin core elements that we took into consideration;

Table 1: Dublin core elements

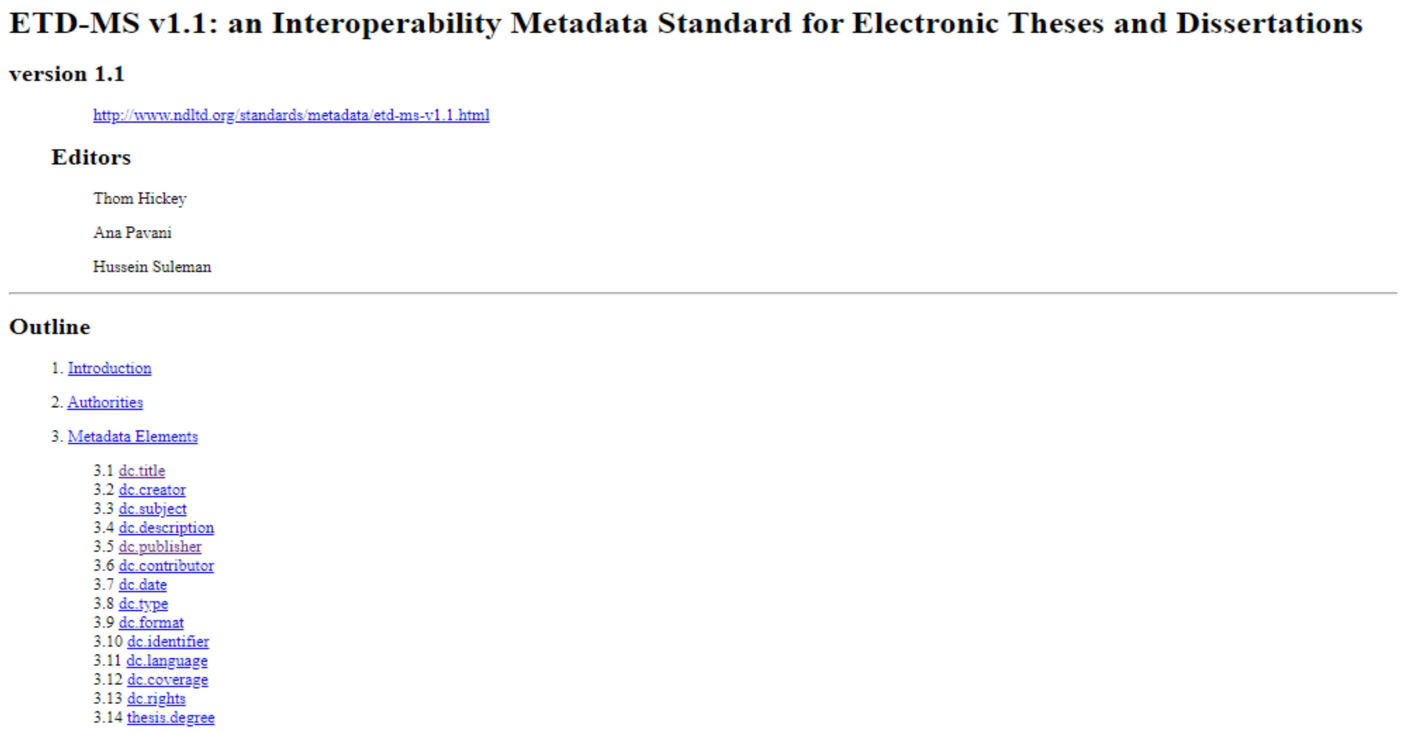
Dublin Core Element	**Definition**
Contributor	An entity responsible for making contributions to the resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Creator	An entity primarily responsible for making the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Description	An account of the resource
Format	The file format, physical medium, or dimensions of the resource
Identifier	An unambiguous reference to the resource within a given context
Language	A language of the resource.
Publisher	An entity responsible for making the resource available
Relation	A related resource
Rights	Information about rights held in and over the resource.
Source	A related resource from which the described resource is derived
Subject	The topic of the resource

Title	A name given to the resource.
Type	The nature or genre of the resource

2.1.2 Metadata in ETD-MS

ETD-MS establishes a standardized set of metadata elements designed for detailing electronic theses or dissertations. Various institutions engaged in handling electronic theses and dissertations have formulated their own standards or modified existing metadata standards[3]. As outlined in the documentation, these metadata standards strive to comprehensively portray details about the author, the work itself, and the contextual backdrop of its production. The objective is to provide valuable information not only to researchers but also to librarians and technical staff responsible for maintaining the electronic version of the work. It is crucial to note, however, that the source document doesn't serve as a substitute for the metadata schemes developed by individual universities or specific environments. Instead, it should be viewed as a reference document aiding in the creation of an effective correlation between local metadata standards and a unified standard for disseminating information regarding ETDs. Figures 2.1.0 and 2.2.0 below illustrate screenshots of the ETD-MS metadata standard, showcasing the prescribed metadata elements.

Below is a sample format of the metadata standard containing all 14 metadata elements, this is according to ETD-MS;



ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations

version 1.1

http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html

Editors

Thom Hickey

Ana Pavani

Hussein Suleman

Outline

1. Introduction
2. Authorities
3. Metadata Elements
   - 3.1 dc.title
   - 3.2 dc.creator
   - 3.3 dc.subject
   - 3.4 dc.description
   - 3.5 dc.publisher
   - 3.6 dc.contributor
   - 3.7 dc.date
   - 3.8 dc.type
   - 3.9 dc.format
   - 3.10 dc.identifier
   - 3.11 dc.language
   - 3.12 dc.coverage
   - 3.13 dc.rights
   - 3.14 thesis.degree

Figure 1.1 ETD-MS metadata elements

The ETD-MS and Dublin core metadata elements are linked by their common objective of establishing standards for the description of electronic theses and dissertations. Each of these standards presents a set of guidelines and elements aimed at ensuring a uniform and thorough representation of metadata. Their connection is rooted in their joint contribution to fostering efficient and standardized metadata practices in both academic and digital library settings.

The ETD Pre-processing Pipeline Portal successfully integrated Dublin Core and ETD-MS elements during its development. A mapping strategy was established for seamless metadata transition, and a standardized representation was enforced. The system included mechanisms for easy metadata transformation and validation, along with user-friendly interfaces and comprehensive documentation for users. The portal was designed to adapt to future standards, and regular quality assurance audits ensured consistent adherence to both Dublin Core and ETD-MS standards. This approach has made the portal a practical solution for efficiently managing electronic theses and dissertations with standardized information.

2.1.3 Current State of ETD Portals and Metadata Quality Enhancement:

The existing literature highlights that while ETD portals have improved access to scholarly works, the issue of metadata quality remains a challenge. Various institutions have taken different approaches to enhance metadata quality. Some have manually curated metadata, while others have explored automated methods, including rule-based systems and machine learning-driven pipelines.

In the context of the Zambia National ETD portal, prior work has resulted in the prototype of an ETD portal designed to archive works from HEIs. However, the need to improve metadata consistency and completeness remains. The proposed thesis addresses this need by introducing a machine learning-driven pre-processing pipeline that aims to address these challenges comprehensively.

Furthermore, the National Electronic Theses and Dissertations (ETD) Portal of South Africa stands as a significant milestone in the realm of academic research and resource accessibility. Designed to be a centralized repository for ETDs across various universities, this portal embodies a multi-tiered architecture, complex yet robust, aimed at harmonizing the landscape of ETD programs and enhancing scholarly interactions.

The NETD Portal is a sophisticated digital platform designed to offer access to a curated collection of Electronic Theses and Dissertations (ETDs) originating from various universities across South Africa. The system operates on a multi-tiered architecture, comprising distinct yet interdependent components: the Harvester, the Repository, and the Web Portal. This architecture synergistically combines Java programming, Java Servlet Web technology, MySQL databases, Lucene search engine, and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to form a cohesive ecosystem for metadata harvesting, storage, and retrieval[4].

The ETD PreProcessing Pipeline Portal aims to address and improve the challenge Metadata enhancement and standardization which is one of the challenges that the NETD portal faces. The ETD PreProcessing Pipeline is an important software tool that will automate data extraction and the standardization process, it will ensure the ETD records are consistently formatted and enriched with more accurate, structured metadata to ultimately enhance discoverability and user experience.

2.1.4 Metadata Harvesting

Metadata harvesting is a process used in digital libraries and information systems to collect metadata, typically from various sources and compile it into a centralized repository. This process enables efficient searching, discovery and access to a wide range of resources, such as scholarly journals, theses and dissertations and many more. [5] Metadata harvesting is commonly employed in the context of digital repositories, institutional repositories, library catalogs and digital libraries. Metadata harvesting relies on a harvesting protocol that standardizes formats for collecting metadata. The Open Archives Initiative Protocol for Metadata (OAI-PMH) is a widely used protocol that enables mass compatibility harvesting of metadata from repositories henceforth it has been used in the development of the Zambia National ETD Preprocessing Pipeline Portal

2.1.5 ETD Aggregation and Accessibility

Aggregating ETDs from diverse HEIs and making them accessible through online platforms provides researchers, scholars, and the broader academic community with an invaluable resource for exploration and research. The Zambia National Electronic Theses and Dissertations (ETD) portal, as described in the thesis, serves as a pivotal platform for the aggregation and dissemination of ETDs from multiple institutions. This portal plays a crucial role in ensuring that scholarly output is easily discoverable and accessible.

However, a common challenge in ETD aggregation is the inherent variability in metadata quality and completeness. Different institutions might use diverse metadata schemas, leading to inconsistencies in formatting and the presence of missing information. Inconsistent metadata not only hampers the discoverability of ETDs but also affects the accuracy of search results and impedes effective cross-referencing. Therefore, efforts to enhance metadata consistency and completeness are essential to realize the full potential of ETD portals.

The ETD Preprocessing Pipeline will consist of records from many institutions much like the NDLTD, of which each of these has a process in place for archiving and distributing ETDs. The NDLTD Union Catalog is an attempt to make these individual collections appear as one seamless digital library to researchers seeking out theses and dissertations

The NDLTD Union Catalog initiative sought to address challenges associated with integrating diverse ETD collections from institutions worldwide[6]. Initially adopting a federated search approach, the project encountered challenges such as varying site availability, evolving search interfaces, and the complexity of merging disparate search results. The transition to the OAI-PMH protocol offered a more consistent and standardized approach to metadata harvesting, thereby simplifying cross-archive searches and facilitating downstream services.
However, challenges persisted in maintaining consistent metadata representation across institutions with disparate practices. The paper introduces a metadata aggregation process as a solution, standardizing metadata to a common schema to ensure consistent search and retrieval experiences. This solution enhances downstream services by providing reliable, structured, and uniform metadata for users and systems that rely on the NDLTD Union Catalog.

CHAPTER THREE

3. Methodology

The data collection techniques to be used will include data extraction from Institutional Repositories of Zambian HEIs. The project team identified a list of Zambian HEIs that offer advanced postgraduate programmes resulting in the publication of theses and dissertation manuscripts. The team extracted metadata from the Institutional Repositories of these HEIs, including author name, title, date of publication, abstract, and keywords. The sampling technique used will be a purposive sampling method, where specific ETDs that met the inclusion criteria were selected.

3.1 Investigate the inconsistencies of metadata from institutional repositories

Metadata was collected from different repositories by the use of the OAI-PMH validator which was used to harvest metadata XML files. Metadata from the University of Zambia was successfully harvested and other metadata files were downloaded and manually analyzed with the use of Microsoft Excel e.g analysis of the consistency of author names such as the casing, format uploading the names'. The harvesting of metadata was done by copying the URL link for different IRs and pasting them onto the OAI-PMH validator, however we only managed to download the XML file from the UNZA and it contained 4789 metadata records. It is at this stage that our research mainly centered on the University of Zambia we then exported the record files as Excel Spreadsheet for data analysis due to challenges in accessing metadata from other IRs.

Data analysis was done on an Excel Spreadsheet to identify inconsistencies in various fields of the associated metadata elements to the UNZA repository and compared with ETD-MS metadata standard.

3.1.1 Challenge

One of the challenges of achieving our first objective was that some repositories such as NIPA, Mulungushi repository, ZAOU were inaccessible due to maintenance and other reasons not known to us. The other challenge was the extraction of metadata from institutional repositories using the oai-pmh validator due to maintenance of the system, the led us to narrow down our research focus to the University of Zambia

3.2 To design and implement the pre-processing pipeline software tool for ETD metadata

In order to design and implement the ETD preprocessing pipeline portal, we drew inspiration from our related works which are the National ETD portal of South Africa and Networked Digital Library of Theses and Dissertations because of their functionality and workflow of the systems.

We conducted expert interviews with stakeholders to chat about requirements with regards to the software design therefore gaining valuable insights as a valuable practice in the software design process for several reasons:

- In-depth understanding of requirements, expert interviews enabled us to gain a profound understanding of the requirements and expectations from individuals who possess deep domain knowledge and expertise. Experts provided us detailed insights into specific aspects of the software design, ensuring that the final product meets our design objectives..

The use of related work and elicited requirements before embarking on our software development project, it's crucial to understand what has been done in the related fields. Analyzing related work provided insights into existing solutions, technologies, and methodologies.

-Inspiration for Innovation: It provides inspiration for innovation by building upon or improving existing solutions.

-Identification of Gaps: Analyzing related work helps identify gaps or shortcomings in existing solutions, guiding your project to address those gaps.

-Eliciting requirements involves systematically gathering, analyzing, and documenting information to understand what stakeholders need from the software.

The benefits of eliciting requirements include;

- Aligning with Stakeholder Expectations: Elicited requirements ensure that the software aligns with the expectations of end-users, clients, and other stakeholders.

-Minimizing Scope Creep: Clearly defined requirements help in minimizing scope creep by providing a well-defined roadmap for development.

-Basis for Testing and Validation: Requirements serve as the basis for testing and validating the final product against what was initially envisioned.

Validation of design decisions because seeking input from experts helps in validating design decisions made during the software development process. Experts can evaluate proposed designs based on their extensive experience, helping to confirm the viability and appropriateness of chosen approaches.By involving experts early on, potential design flaws or issues can be identified and addressed before they become critical problems.
-appropriateness of chosen approaches.

-Alignment with industry best practices because experts often have a deep understanding of industry best practices and standards. Therefore, expert interviews help ensure that the software design aligns with established best practices, contributing to the overall quality and reliability of the product. On the other hand, experts can guide the design process to ensure compliance with industry standards and regulations.

3.2.1 Softwares used

Microsoft Excel is a spreadsheet tool that offers a variety of features for data analysis. The following are some of the features we used for analysis of data;

- Data sorting and filtering arrange and organize data by sorting columns or applying filters to focus on specific information.
- Data validation to set rules and constraints to ensure data accuracy and integrity

Google Meet is a video conferencing and collaboration tool that offers various features for online meetings and communication. The following are the features we utilized on Google Meet;

- Online meetings and video conferencing were conducted with supervisors, colleagues, stakeholders, and team members for video and audio calls, screen sharing, and real-time collaboration.

3.3 To demonstrate the complete consistency of Pre-processed ETD metadata .

The intention is to show that the metadata of electronically processed theses and dissertations (ETDs) has been handled in a manner that ensures thorough uniformity. Pre-processed ETD metadata" refers to the information describing the data associated with electronic theses and dissertations before any further processing or manipulation. To ensure complete consistency of the metadata in a way that ensures coherence, uniformity and reliability. All the metadata elements or attributes are expected to be in agreement and conform to a standardized format or set of criteria.

We carefully created rules for organizing information in Electronic Theses and Dissertations (ETD), covering important details like author info, title, and summary. It was crucial for us to match these rules smoothly with what our school needed, making sure everything fits well into our academic system. To do this precisely, we used established guidelines, borrowing from the strong ETD-MS (Electronic Theses and Dissertations Metadata Standard) and the well-known Dublin Core system for organizing information[8]. By sticking to these worldwide-accepted rules, we not only kept the information's quality and consistency but also showed our commitment to doing well in academics and working together with other schools. This smart mix of ETD-MS and Dublin Core mirrors our promise to have a clear and easy way of organizing information that benefits writers, researchers, and everyone in the academic community.

3.4 To have a unified interface for the presentation of the ETDs

In an attempt to achieve our last objective which is to have a unified interface for the presentation of Electronic Theses and Dissertations, we conducted expert interviews with the stakeholders directly involved and affected by the ETD presentation platform. These interviews we conducted both online and physically.

To identify common themes, preferences, and requirements, we utilized qualitative data analysis techniques, such as thematic coding, to categorize and interpret responses and identify recurring patterns and key insights to inform the design and functionality of the unified interface[10].

CHAPTER 4

4. Results and Discussion

4.1.1 Introduction

Tools Utilized for Evaluation:
To conduct this analysis, we employed a combination of manual inspection and automated tools. Manual inspection allowed us to discern the nuances of casing variations, while automated tools facilitated the identification of specific patterns, such as Title Casing/ Sentence case. The use of both approaches ensured a comprehensive evaluation of the title field metadata.

4.1.2 Metadata Analysis in the UNZA Repository

Ensuring the quality of metadata is crucial for the discoverability, usability, and overall integrity of digital repositories. In this comprehensive analysis, we delve into the inconsistencies identified within specific metadata fields in the University of Zambia's (UNZA) repository. Our primary objective is to pinpoint these discrepancies and propose actionable recommendations to enhance the overall quality of the repository's metadata.

4.1.3 Title Field Metadata Analysis

The chart below reveals that there are inconsistencies in the title field metadata within the UNZA repository. These inconsistencies include the absence of titles, titles in uppercase, and titles in CamelCasing. Addressing these inconsistencies is essential to improve the overall quality and usability of the repository's metadata. Actions such as adding titles to the theses with no titles, converting uppercase titles to title case, and standardizing the casing of theses with CamelCasing can help enhance the repository's metadata quality.
Title field metadata analysis revealed the following inconsistencies:

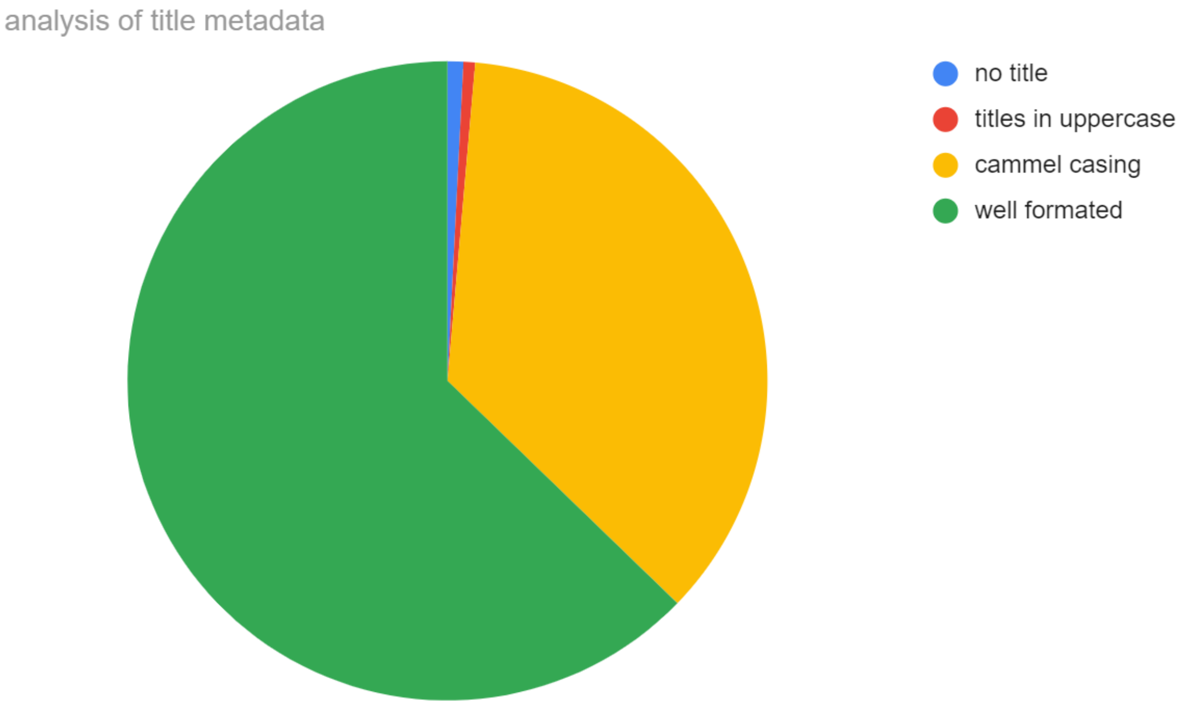


Figure 2.1 analysis of title metadata

Thesis with No Title (54): 0.8%

It's concerning that there are 54 theses with no titles. This might make it difficult for users to understand the content and purpose of these theses.

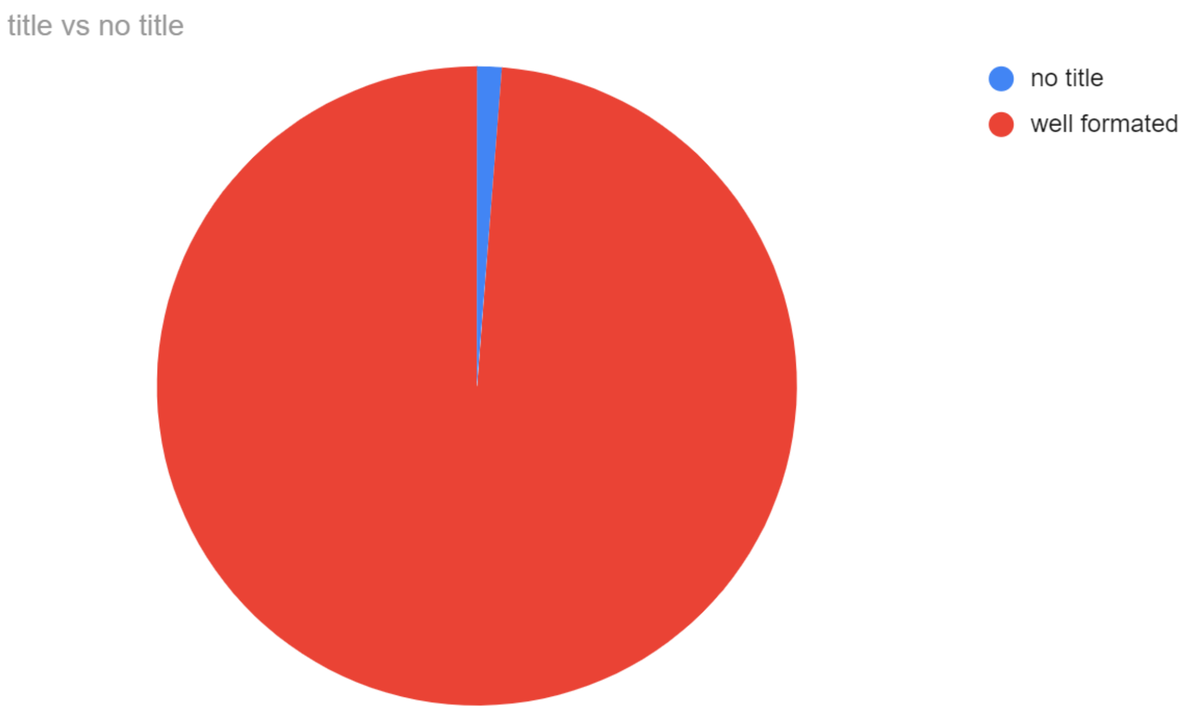


Figure: 2.2 Thesis with no title

Impact: The absence of titles can hinder users' understanding of thesis content.
Recommendation: Implement a process to add appropriate titles to these theses.

Thesis with Titles in Uppercase (40): 0.6%

Having 40 theses with titles in uppercase suggests a consistency issue. Title case or sentence case is typically used for titles to improve readability and follow standard formatting conventions. These theses should be converted to the correct casing.

Impact: Uppercase titles deviate from standard formatting conventions.
Recommendation: Convert these titles to title case for readability.

Thesis with CamelCasing (2,431): 35.9%

The presence of 2,431 theses with CamelCasing indicates a significant inconsistency. While CamelCasing is a valid formatting choice for certain contexts, it might not be the best choice for thesis titles, as it can affect readability.

Impact: CamelCasing may affect readability and consistency.
Recommendation: Standardize casing to title case for all titles.

Well-Formatted Thesis (4,256): 62.8%

Having 4,256 well-formatted theses is a positive sign. These theses are likely following the standard title formatting rules, which is crucial for metadata consistency and user experience.

Impact: These well-formatted theses adhere to metadata standards.
Recommendation: Continue to maintain metadata consistency and quality.
Institutional Names Metadata Analysis

4.1.4 The institutional names metadata analysis revealed the following inconsistencies:

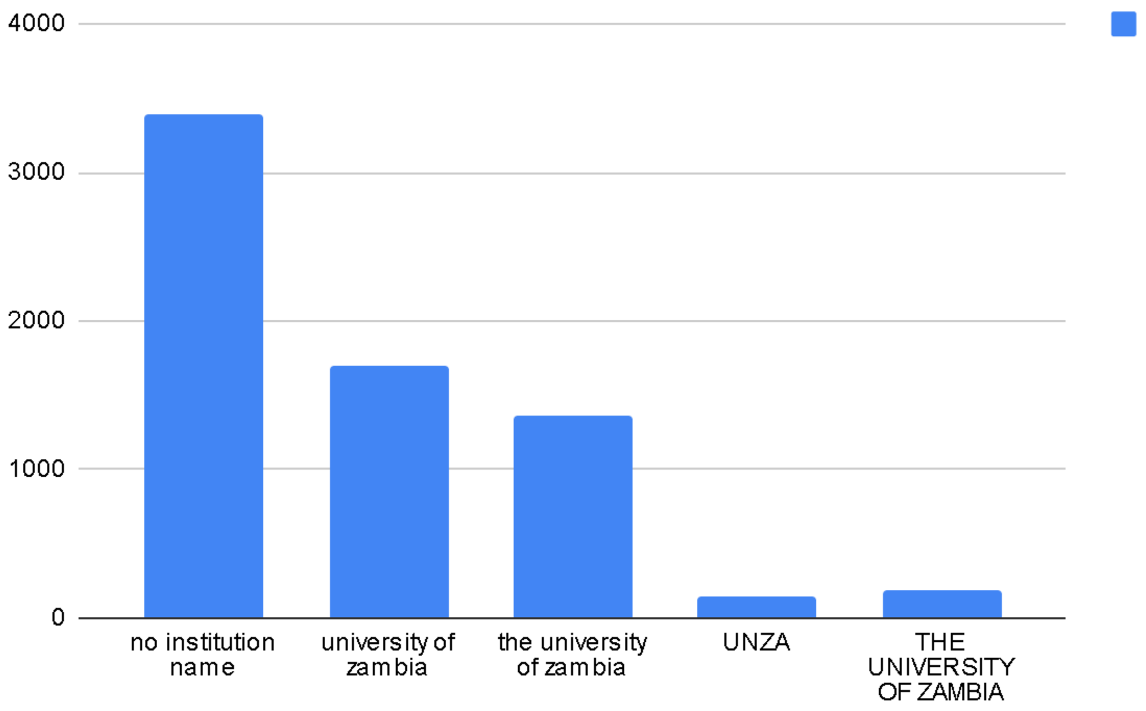
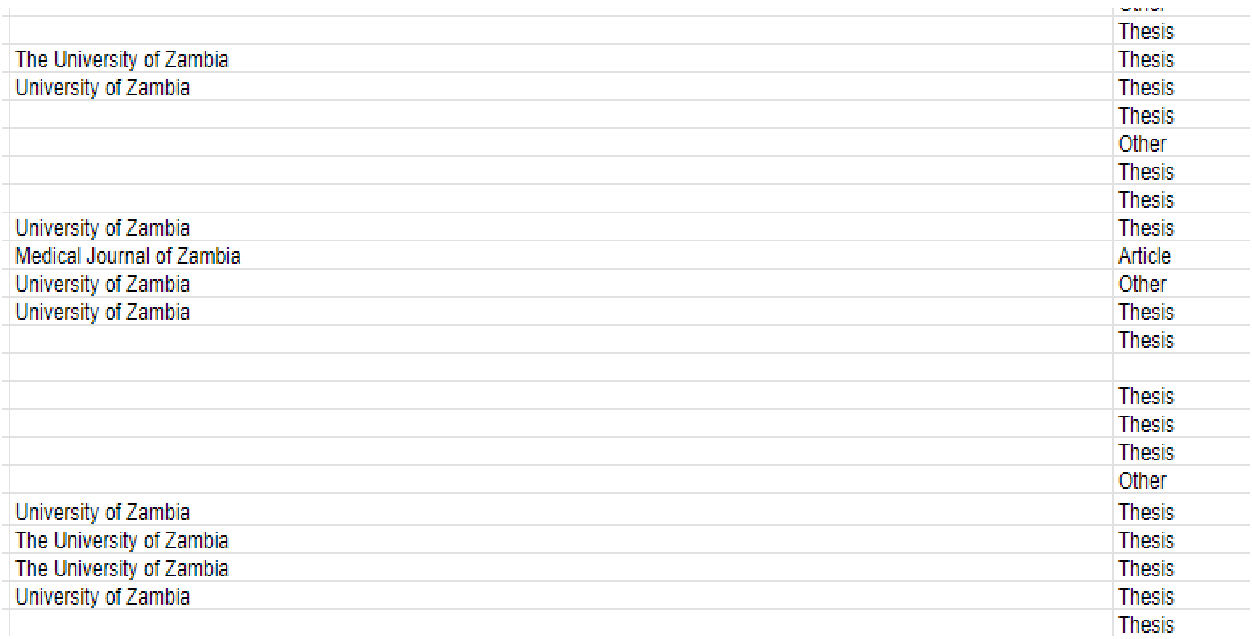


Figure 2.3 Inconsistencies in institutional name

No Institution Name (3,391):

The presence of 3,391 entries with no institution name is a significant concern. An institution's name is a fundamental part of metadata and is critical for identification and organization.



	Thesis
The University of Zambia	Thesis
University of Zambia	Thesis
	Thesis
	Other
	Thesis
	Thesis
University of Zambia	Thesis
Medical Journal of Zambia	Article
University of Zambia	Other
University of Zambia	Thesis
	Thesis
	Thesis
	Thesis
	Thesis
	Other
University of Zambia	Thesis
The University of Zambia	Thesis
The University of Zambia	Thesis
University of Zambia	Thesis
	Thesis

Figure 2.4 Missing institutional names

Impact: The absence of institution names can result in loss of context.
Recommendation: Identify and add the correct institution names for these entries.

Not only that i was found that the institutional names where written in different variations

University of Zambia (1,695):

The term "University of Zambia" (in title case) is the correct institution name, and it's good to see a substantial number of entries using this. However, there are other variations that should be standardized for consistency.

Impact: This is the correct institution name, but other variations exist.
Recommendation: Standardize all instances to "University of Zambia" in title case.

The University of Zambia (1,357):

"The University of Zambia" is essentially the same institution as "University of Zambia" but with the definite article "The." Standardizing this to "University of Zambia" (title case) may help improve consistency.

Impact: Similar to "University of Zambia" but with variations.
Recommendation: Standardize to "University of Zambia" for consistency.

UNZA (150):

"UNZA" is an acronym for the University of Zambia. It's a common abbreviation for institutions, but it should be consistently defined as "University of Zambia" for clarity and consistency.

Impact: "UNZA" is an acronym for the University of Zambia.
Recommendation: Consistently define as "University of Zambia" for clarity.

THE UNIVERSITY OF ZAMBIA (188):

Entries in all uppercase, such as "THE UNIVERSITY OF ZAMBIA," are inconsistent with standard title case formatting. It's important to standardize these entries to "University of Zambia" for uniformity.

Impact: All uppercase entries deviate from standard formatting.
Recommendation: Standardize to "University of Zambia" in title case.

Type Field Metadata Analysis

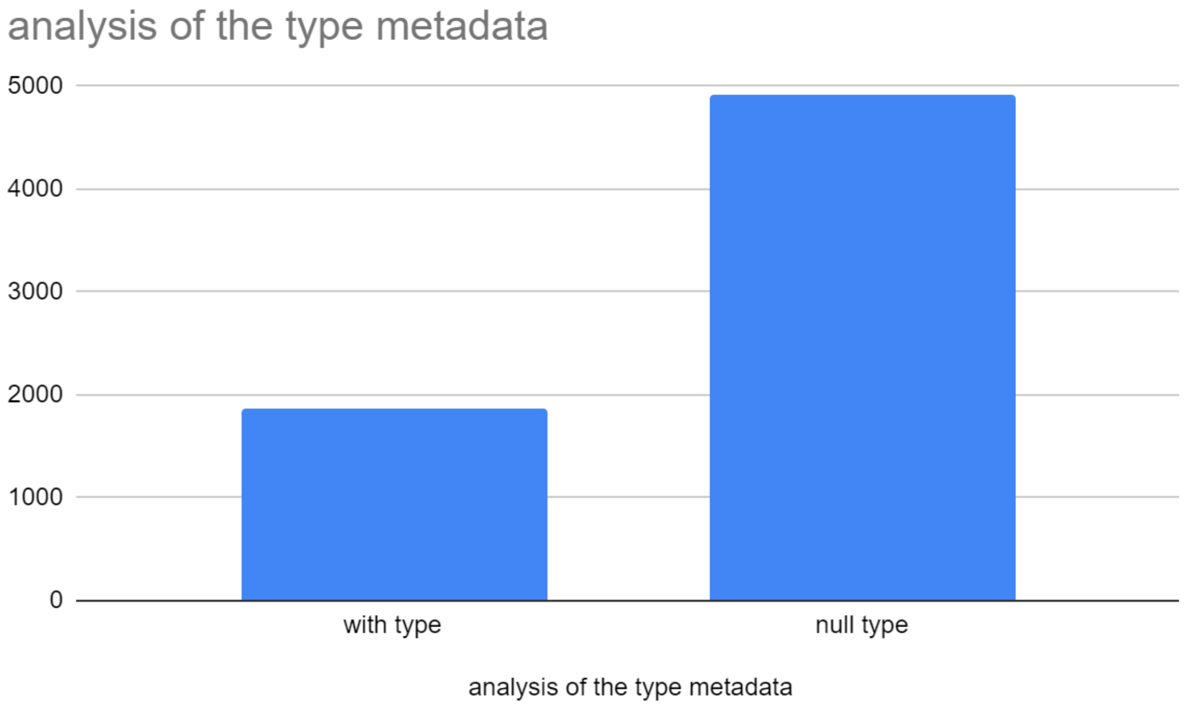


Figure 2.5 metadata Type analysis

4.1.5 Thesis Type Metadata:

1860 ETDs have type metadata.
4927 ETDs have no type metadata.

In the UNZA repository, approximately 27.34% of the ETDs have type metadata, while about 72.66% of the ETDs lack this essential metadata. This inconsistency highlights the importance of a pre-processing pipeline and machine learning models to augment missing metadata and ensure comprehensive descriptions of ETDs in the repository.

Justification for Building the Preprocessing Pipeline The analysis of title and institutional names metadata inconsistencies underscores the urgent need for a preprocessing pipeline software tool within the UNZA repository. Here's the justification:

Enhanced Metadata Quality: Inconsistencies in metadata, as demonstrated, affect the usability and discoverability of digital resources. A preprocessing pipeline can automate processes to improve metadata quality by adding missing information and standardizing formattin
g.

Improved User Experience: Researchers, librarians, and administrators rely on accurate and consistent metadata to effectively manage and access ETDs. The preprocessing pipeline will ensure that titles and institutional names adhere to standards, leading to a more user-friendly experience.

Efficiency and Automation: Manual correction of thousands of metadata entries is time-consuming and prone to errors. The preprocessing pipeline can automate these tasks, saving time and resources for more critical activities.

Consistency and Compliance: Standardizing metadata ensures compliance with metadata standards and best practices. This, in turn, supports interoperability and data exchange with other systems.

Data Enrichment: The pipeline can enrich metadata by identifying missing titles and institution names, thereby enhancing the overall information available to users.

In conclusion, the analysis of metadata inconsistencies in the UNZA repository highlights the urgent need for a preprocessing pipeline software tool. Such a tool will not only improve metadata quality but also enhance the user experience, streamline processes, ensure compliance, and enrich the repository's data. Building the preprocessing pipeline is a crucial step toward maintaining a high-quality digital repository.

4.2 Risk Analysis

In the context of your project, "risk analysis" refers to the systematic process of identifying, assessing, and managing potential risks that could impact the successful development and implementation of the ETD (Electronic Theses and Dissertations) Preprocessing Pipeline. Risk analysis is a crucial component of project management, due to the fact that it helped our team to anticipate challenges and uncertainties and allows us to develop effective strategies for mitigating or managing these risks.

4.2.1 Risk Identification

Machine Learning Model Performance:

- Risk: The machine learning models may not perform as expected, leading to inaccuracies in metadata augmentation.

- Mitigation:
  Conduct extensive testing and validation of machine learning models before deployment.
  Implement a feedback loop for continuous improvement based on real-world usage and feedback from librarians and users.

Integration Issues with External Systems:

- Risk: Difficulties may arise when integrating the ETD Preprocessing Pipeline with external library catalog systems.
- Mitigation:
  Collaborate closely with external system owners to understand integration requirements.
  Conduct thorough testing of integration points and implement fallback mechanisms in case of integration failures.

Limited Resources and Skilled Personnel:

- Risk: Shortage of skilled developers or analysts may slow down the development process.
- Mitigation:
  Assess resource needs early and consider outsourcing if necessary.
  Invest in training programs for existing team members to enhance their skills.
  Build a diverse and cross-functional team to mitigate the risk of dependencies on specific individuals.

Scalability Issues:

- Risk: The software may face challenges in handling a growing volume of ETDs and users.
- Mitigation:
  Design the software architecture with scalability in mind.
  Regularly monitor system performance and conduct scalability testing.
  Plan for infrastructure scaling and optimization based on usage patterns.

Lack of Documentation:

- Risk: Inadequate documentation may lead to challenges in maintenance and troubleshooting.
- Mitigation:
  Document the software architecture, codebase, and deployment processes thoroughly.
  Maintain up-to-date user and administrator documentation.
  Implement a documentation review process to ensure accuracy and completeness.

4.2.2 RISK PLAN

Table 2: Risk analysis and plan

	Risk Category	Risk Description	Likelihood	Impact	Overall Risk Level	Mitigation Strategies
1	Machine Learning Model Performance	The models may not perform as expected, leading to inaccuracies in metadata augmentation.	Medium	High	High	- Conduct extensive testing and validation before deployment. - Implement a feedback loop for continuous improvement based on real-world usage and feedback.
2	Integration Issues with External Systems	Difficulties in integrating with external library catalog systems.	Medium	Medium	Moderate	Collaborate closely with external system owners. - Conduct thorough testing of integration points. - Implement fallback mechanisms

						for integration failures.
3	Limited Resources and Skilled Personnel	Shortage of skilled developers or analysts may slow down development.	Medium	High	moderate	- Assess resource needs early and consider outsourcing if necessary. - Invest in training programs for existing team members. - Build a diverse and cross-functional team.
4	Scalability Issues	Challenges in handling a growing volume of ETDs and users	Low	High	Medium	- Design the software architecture with scalability in mind. - Regularly monitor system performance and conduct scalability testing. - Plan for infrastructure scaling based on usage patterns.
5	Lack of Documentation	Inadequate documentation may lead to challenges in	Low		Moderate	- Document the software architecture,

		maintenance and troubleshooting.				codebase, and deployment processes thoroughly. - Maintain up-to-date user and administrator documentation. - Implement a documentation review process.
6	Data Security and Privacy	Unauthorized access to sensitive ETD data could compromise data integrity and user privacy.	Low		High	- Implement robust authentication mechanisms. - Conduct regular security audits and ensure compliance with data protection regulations. - Educate users on security best practices.
7	Insufficient User Training and Adoption	Users may struggle to adapt due to inadequate training resources.	Medium		Moderate	- Develop comprehensive user training materials. - Conduct training

						sessions and establish a helpdesk. - Provide ongoing support and documentation.

4.2.3 Resources

The following are the resources that are needed that to successfully develop the project:

Technical Resources:

To successfully develop the ETD Preprocessing Pipeline, a suite of technical resources is essential. As a team we organized some laptops equipped with IDEs (Integrated Development Environments), in our case we chose Visual Studio Code. Additionally, version control systems like Git will be utilized for collaborative coding and version management.

Learning and Knowledge Resources:

Access to comprehensive course materials, textbooks, and online resources related to software development, machine learning, and database management is crucial. Furthermore, leveraging online courses and platforms such as Coursera or Udemy will aid in acquiring specific skills related to programming and machine learning. Well-documented resources, tutorials, and guides will be essential for utilizing ML libraries and programming languages effectively.

Infrastructure and Software:

We will also make use of Cloud computing platforms like AWS, Linode or Google Cloud for hosting and deployment purposes. Databases we chose MySQL. MySOL will facilitate efficient management of ETD metadata. Availability of machine learning libraries including TensorFlow and PyTorch is imperative for the development of ML models.

Collaboration and Communication:

Efficient team collaboration will rely on communication platforms like Slack or Microsoft Teams. Project management tools such as Trello or Jira will assist in task management and progress tracking, ensuring seamless coordination among team members.

Training and Support:

Attending workshops or seminars related to software development and machine learning will enhance skill sets. Access to mentors, advisors, or faculty members for guidance and support throughout the project will be beneficial.

Data and Information:

Access to sample datasets specifically related to Electronic Theses and Dissertations (ETDs) will aid in testing and training ML models.

Accessibility and Environment:

Reliable internet services and a suitable workspace conducive to coding and project development are prerequisites for uninterrupted progress.

Time and Commitment:

Dedication of committed time and effort from team members towards the project's development, meetings, and collaboration is imperative for successful execution.

These resources collectively form the backbone of the project development process, ensuring a comprehensive approach towards the successful realization of the ETD Preprocessing Pipeline.

4.2.4 Deliverables and Milestones

The following deliverables are expected to be produced after successful completion of the project:

- **Project website**

We designed and deployed a project website, a project which hosts our project documentation and other deliverables.

- **Project proposal document**

We gave a comprehensive project plan outlining the scope, objectives, resources and milestones and timelines such as the Gantt Chart, work breakdown structure and activity network diagram.

- **Software tool**

We developed a fully functional software application tool the ETD pre-processing pipeline

meeting specified requirements such as the ability to consistently format metadata and a unified interface.

- **Final report**

  The Final Project report was successfully compiled and delivered.

CHAPTER FIVE

5. System Design and Implementation

5.1 Project Overview

The primary objective of this project is to develop a robust preprocessing pipeline software designed to enhance and standardize metadata that arrives in an unformatted state. Our focus lies in systematically processing metadata generated from various sources, particularly addressing instances where metadata lacks consistency or structure.

Objectives:

Metadata Preprocessing: The central aim is to preprocess unstructured or inconsistently formatted metadata from different sources, ensuring a unified and standardized format.

Consistency Enhancement: To augment missing or incomplete metadata using machine learning models, thereby improving overall data consistency.

Scope

The project encompasses the development of a sophisticated software solution that not only preprocesses metadata but also integrates machine learning models to augment and standardize missing or inconsistent metadata. It covers the processing of metadata from various Higher Education Institution (HEI) sources within Zambia.

Significance and Impact:

This project's significance lies in its capacity to revolutionize the handling of unstructured metadata within academic repositories. By creating a standardized and enriched metadata repository, it aims to significantly ease the accessibility and retrieval of ETDs for academic researchers, students, and library administrators. Moreover, the streamlined processing of metadata ensures a more efficient and coherent national archive of academic work, contributing to the advancement of scholarly research and knowledge dissemination.

5.1.1 System Requirements Analysis

Gathering and Analyzing System Requirements:

The process of system requirements gathering involved a comprehensive approach. Initially, extensive stakeholder consultations were conducted with librarians and end-users to understand their needs and challenges regarding metadata formatting. Additionally, in-depth reviews of existing institutional repositories and metadata standards were conducted to comprehend the current landscape and potential gaps.

Functional Requirements:

Metadata Preprocessing: Ability to preprocess metadata from various sources.

Machine Learning Augmentation: Implementing ML models to augment and standardize missing metadata.
User Interface: Developing an intuitive and user-friendly interface for metadata management.
Search and Retrieval: Enabling efficient search and retrieval functionalities for ETDs.

Non-functional Requirements:

Performance: Ensuring the system operates efficiently even with large volumes of metadata.
Scalability: Designing the system to handle increasing data and user loads.
Usability: Creating an interface that's easy to navigate and understand for users of varying

technical backgrounds

.

Prioritization and Documentation:

Requirements prioritization was based on several factors including stakeholder feedback, impact on system functionality, and feasibility within project timelines. A collaborative effort involving the project team and stakeholders led to the creation of a prioritized requirements list. These requirements were meticulously documented in a structured manner, ensuring clear delineation between functional and non-functional aspects. Each requirement was detailed with its description, rationale, and its impact on the overall system architecture and user experience.

Methodology:

The Agile methodology, specifically Scrum, was employed to facilitate requirements prioritization and documentation. Regular sprint planning sessions allowed for continuous refinement and reassessment of requirements based on changing stakeholder needs and emerging insights from development activities. This iterative approach ensured that evolving project demands and stakeholder expectations were accommodated throughout the development lifecycle.

5.1.2 Institutional Repository Software Tools

In the pursuit of establishing a robust preprocessing pipeline for our software development initiative, the selection of Institutional Repository Software Tools is a pivotal step. This section provides an in-depth exploration of the chosen tools—DSpace, OAI-PMH, and Linode—highlighting their significance, features, and how they synergistically contribute to the efficiency of our software architecture.

DSpace

Overview:

DSpace stands as a cornerstone in the realm of institutional repositories, offering a comprehensive and highly customizable solution for managing digital assets[10]. Its open-source nature, coupled with a rich set of features, makes DSpace an ideal choice for our preprocessing pipeline.

Key Features:

Scalability: DSpace is renowned for its ability to scale effortlessly, accommodating the increasing volume and diversity of digital content within institutional repositories.

Community Support: The vibrant and active DSpace community ensures ongoing development, support, and a wealth of plugins, contributing to the adaptability of the software to our specific needs.

Metadata Management: Robust metadata capabilities empower administrators to meticulously categorize, organize, and retrieve resources, enhancing the overall discoverability of content.

OAI-PMH

Overview:

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) plays a pivotal role in ensuring interoperability and efficient sharing of metadata across diverse repositories. Integrating OAI-PMH into our software architecture enhances the accessibility and dissemination of digital assets.

Key Features:

Interoperability: OAI-PMH facilitates seamless interoperability between repositories, allowing for the standardized exchange of metadata, thereby fostering a connected and collaborative digital ecosystem.

Metadata Harvesting: The protocol's ability to harvest metadata from disparate repositories ensures a unified approach to information retrieval, promoting a holistic view of available resources.

Linode

Overview:

The inclusion of Linode in our chosen tools is strategic. Linode provides virtual cloud services on which we host our server, it is the infrastructure that underpins the entire preprocessing pipeline ensuring reliability and performance.

Key Features:

Scalable Hosting: Linode's cloud infrastructure offers scalable hosting solutions, providing the necessary resources to support the growing demands of institutional repositories.

Reliability: Linode's reputation for reliability and uptime ensures the continuous availability of digital assets, contributing to a seamless end-user experience.

The careful selection of DSpace, OAI-PMH, and Linode reflects a holistic approach to building an institutional repository and preprocessing pipeline. These tools collectively address scalability, interoperability, metadata management, and infrastructure reliability, aligning with our overarching goals. The subsequent sections of this report will delve into the integration of these tools, showcasing their specific functionalities within our software development framework.

5.1.3 Workflow Overview

Metadata Ingestion:

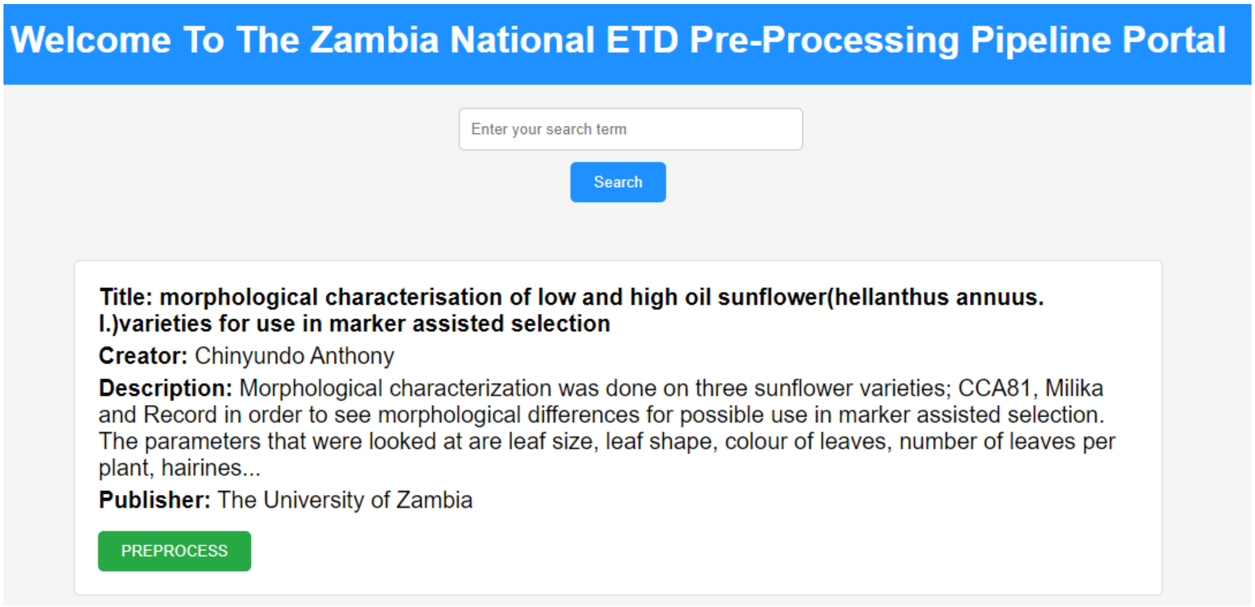
Raw metadata from different Higher Education Institutions (HEIs) and repositories is ingested into the system. This unstructured or inconsistently formatted metadata arrives from various sources and formats.

Metadata Preprocessing:

The metadata undergoes a preprocessing phase where it's validated, cleaned, and standardized. Algorithms within the system process this data, ensuring a unified and consistent format across all entries.

Database Management:

Processed metadata and associated ETDs are stored and managed within a MySQL database. The system ensures efficient storage and retrieval of this information.



Welcome To The Zambia National ETD Pre-Processing Pipeline Portal

Title: morphological characterisation of low and high oil sunflower(hellanthus annuus. l.)varieties for use in marker assisted selection

Creator: Chinyundo Anthony

Description: Morphological characterization was done on three sunflower varieties; CCA81, Milika and Record in order to see morphological differences for possible use in marker assisted selection. The parameters that were looked at are leaf size, leaf shape, colour of leaves, number of leaves per plant, hairines...

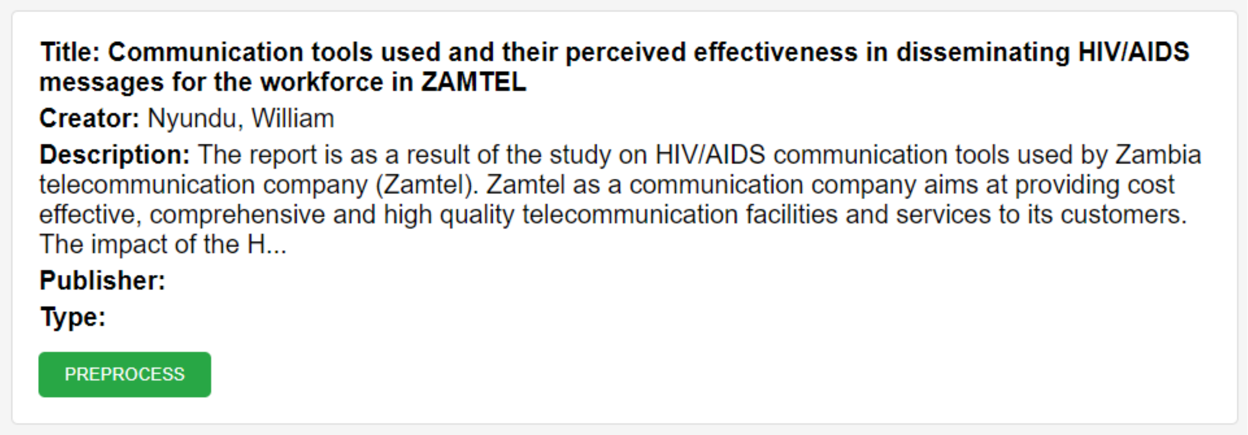
Publisher: The University of Zambia

Figure 3.1 ETD preprocessing pipeline interface

Figure Description:

Figure 1 displays the system's user interface with a prominently displayed "Search" button.

The presence of the "Search" button indicates that users can initiate a search process within the system for a specific Electronic Thesis or Dissertation (ETD). This feature allows users to enter search queries or parameters to locate particular documents or research papers within the system's database.



Title: Communication tools used and their perceived effectiveness in disseminating HIV/AIDS messages for the workforce in ZAMTEL

Creator: Nyundu, William

Description: The report is as a result of the study on HIV/AIDS communication tools used by Zambia telecommunication company (Zamtel). Zamtel as a communication company aims at providing cost effective, comprehensive and high quality telecommunication facilities and services to its customers. The impact of the H...

Publisher:

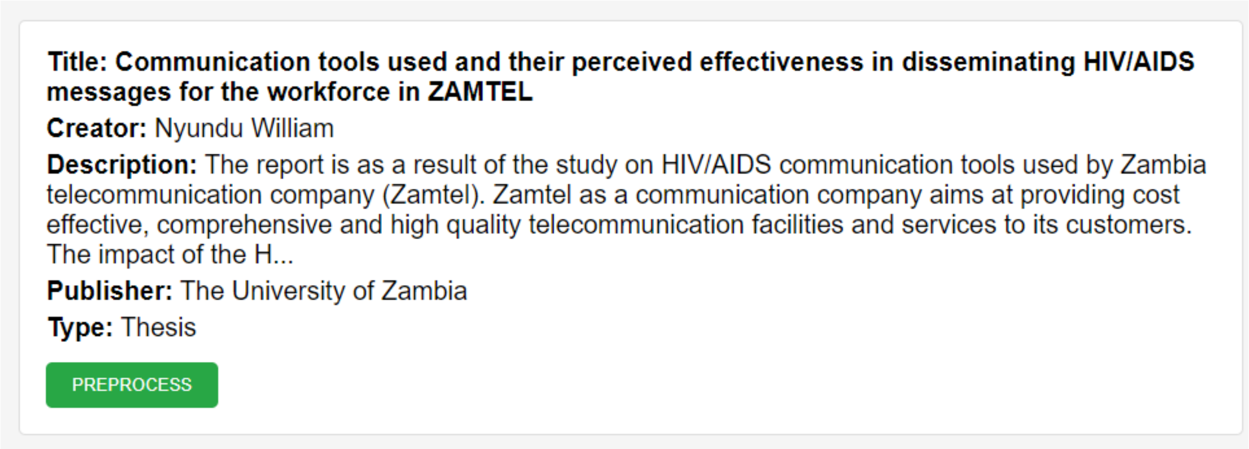
Type:

Figure 3.2 ETD Metadata record before preprocessing

Figure Description:

Figure 2 showcases the output or results obtained after performing a search for a particular ETD.

Upon conducting a search, the system generates and displays the search results, presenting a list of documents matching the entered criteria. If, upon reviewing the results, a user identifies a document with incorrectly formatted metadata or incomplete information, the interface provides an additional functionality—a "Preprocess" button.



Title: Communication tools used and their perceived effectiveness in disseminating HIV/AIDS messages for the workforce in ZAMTEL

Creator: Nyundu William

Description: The report is as a result of the study on HIV/AIDS communication tools used by Zambia telecommunication company (Zamtel). Zamtel as a communication company aims at providing cost effective, comprehensive and high quality telecommunication facilities and services to its customers. The impact of the H...

Publisher: The University of Zambia

Type: Thesis

Figure 3.3 ETD metadata after preprocessing

Figure Description:

Figure 3 exhibits the displayed output of a document after the preprocessing operation.

Upon clicking the "Preprocess" button for a specific document with incorrectly formatted metadata, the system initiates a process to rectify or format the metadata correctly. Figure 3 portrays the transformed output of the document with properly formatted metadata. This view showcases the updated and correctly structured metadata, making it more accessible and usable for users.

5.1.4 Summary:

Figure 1: Indicates the availability of a search feature.

Figure 2: Depicts search results and the "Preprocess" option for correcting metadata.

Figure 3: Demonstrates the corrected and properly formatted metadata output after the preprocessing operation.

These figures collectively illustrate the search functionality, output presentation, and the system's capability to preprocess and rectify incorrectly formatted metadata, ensuring enhanced usability and accuracy of information within the system.

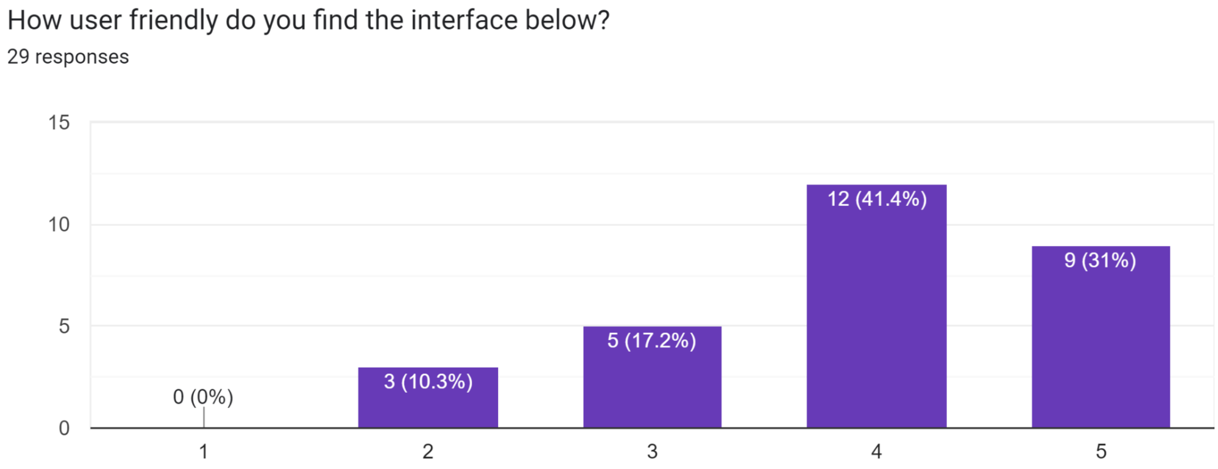
CHAPTER SIX

6. System Evaluation

During the evaluation phase, we facilitated an interactive experience for stakeholders, enabling them to interact with the system then afterwards they were provided with a questionnaire to fill in based on their user experience. This approach was aimed at obtaining valuable feedback to measure the system's performance and functionality, the insights gained from this process were crucial in further refining and improving the system to meet the needs and expectations of the users.

Below is a figure depicting responses of how user-friendly and interactive the users found the system to be;

- 31% of the users found the system to be very user friendly and 41.% of the users found the system to be user friendly which is reasonably good feedback for evaluation purposes.
- However, the reaming 28% found the software to be fairly friendly which set the basis for the improvements and adjustments made



How user friendly do you find the interface below?

29 responses

Figure 4.1 bar chart showing the responses of how user friendly the users found the system

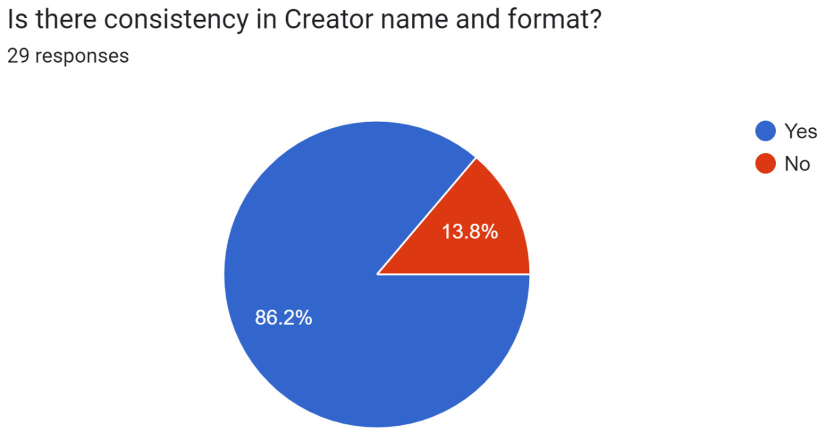


Figure 4.2 Pie chart showing responses of consistency in Creator name and format

One of the critical queries posed was regarding the consistency in the Creator name and format within the metadata. The responses obtained after stakeholders interacted with the system were as follows:

Question: Is there consistency in the Creator name and format?

Responses:

86% of stakeholders responded affirmatively, acknowledging consistency in the Creator name and format within the metadata.

15% of stakeholders indicated a lack of consistency in the Creator name and format.

Evaluation Insights:

The overwhelmingly positive response from 86% of stakeholders signifies a commendable achievement in ensuring consistency within the Creator name and format. This high percentage underscores the system's effectiveness in standardizing and maintaining a coherent structure for the Creator's information across various metadata entries.

However, the 15% minority highlighting inconsistency in the Creator name and format indicates a potential area for further refinement. These responses spotlight the necessity for continued attention to detail and improvement, aiming to address inconsistencies and enhance the overall uniformity in metadata presentation.

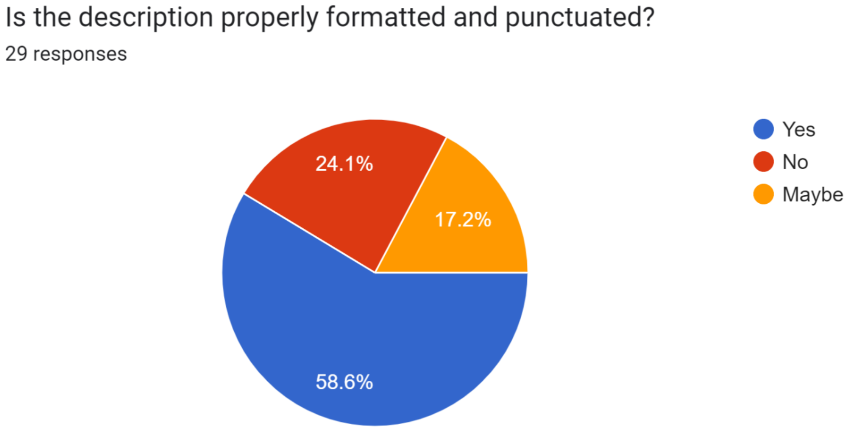


Figure 4.3 Pie chart showing responses of consistency in the description

In the system evaluation, stakeholders were surveyed on the adequacy of the description formatting and punctuation within the metadata. The responses, obtained after user interactions, presented diverse perspectives among different user groups:

Question: Is the description properly formatted and punctuated within the metadata?

Responses:

24% of respondents acknowledged the description as properly formatted and punctuated.

17% of participants indicated that the description lacked proper formatting and punctuation.

**"Maybe" responses from a majority of students were recorded. Due to the extension of the evaluation to students for broader feedback, a conclusive response was not achievable.

Evaluation Insights:

The distinct responses indicate a varied perception regarding the adequacy of description formatting and punctuation within the metadata. The 24% affirmation and 17% critique portray a disparity in stakeholder viewpoints, highlighting the need for further examination and potential enhancements.

The "maybe" responses from a significant portion of the student participants present an inconclusive stance. This outcome underscores the complexity of the evaluation, where diverse user groups might have differing levels of familiarity or expertise in assessing metadata formatting and punctuation.

In conclusion The system evaluation conducted to assess various facets of the preprocessing pipeline software elicited valuable insights and feedback from stakeholders. The culmination of these assessments portrays both commendable achievements and areas warranting further attention and improvement.

Key Findings:

Metadata Consistency: The affirmative response from 86% of stakeholders regarding the consistency in Creator name and format demonstrates the system's success in maintaining uniformity across metadata entries.

Description Formatting: Diverse opinions emerged regarding the adequacy of description formatting and punctuation. While 24% acknowledged proper formatting, 17% expressed concerns, indicating scope for enhancement.

Diverse User Perspectives: The inclusion of students in the evaluation expanded the user spectrum. The variance in responses, particularly the inconclusive "maybe" responses, highlights the necessity for broader engagement to attain comprehensive evaluation insights.

Implications:

Strengths and Achievements: The system's ability to ensure metadata consistency, as validated by the majority response, stands as a significant success contributing to enhanced data uniformity.

Areas for Enhancement: The varied responses regarding description formatting emphasize the need for continuous improvement, guided by stakeholder feedback, to address concerns and refine metadata quality.

Recommendations:

Continuous Refinement: Encourage ongoing refinement cycles, leveraging stakeholder feedback, to address identified areas for improvement, ensuring continual enhancements in metadata quality and system usability.

Diverse Engagement: Broaden engagement with diverse user groups to gather comprehensive feedback, ensuring inclusive evaluations that represent varied user perspectives.

Overall Reflection:

The system evaluation serves as a pivotal guidepost, acknowledging achievements and outlining paths for continual growth. The feedback obtained remains instrumental in steering future developments, fostering a system that meets evolving user needs, and consistently delivers high-quality metadata management solutions.

7. Conclusion

The improvement of the accessibility of ETDs generated by Zambian HEIs that contain consistently formatted metadata and augmented missing metadata is a very keen aspect that needs to be addressed to improve the searchability of Electronic Theses and Dissertations published all the Zambian Higher Education Institutions. It has been an illuminating and enriching experience to develop the Zambia National ETD Preprocessing Pipeline portal that will contribute significantly to the Zambian Higher Educational system

Our primary aim revolved around crafting a robust preprocessing pipeline, intending to streamline metadata processing from various Higher Education Institutions (HEIs) within Zambia.

The project's impact echoes in the realm of academia and research facilitation. Our system's ability to achieve metadata consistency, as lauded by 86% of stakeholders, stands as a testament to its effectiveness in harmonizing Creator name formats, contributing significantly to a more unified data repository in accordance with the ETD-MS and Dublin Core standards.

The journey wasn't without its challenges. Varied feedback on description formatting, where 24% acknowledged adequacy and 17% expressed concerns, illuminated the need for perpetual refinement and user-centric enhancements. This experience underscores the importance of continuous improvement guided by stakeholder input.

Our extended engagement with students underscored the diverse perspectives within our user base. The inconclusive "maybe" responses emphasize the importance of embracing diverse viewpoints and fostering an inclusive environment for user feedback and engagement which is the basis for the software’s future advancements .

Recommendations and Future Prospects:

Moving forward, a commitment to continuous refinement remains our beacon. By harnessing stakeholder insights and broadening user engagement, our system can evolve as a dynamic and user-responsive solution, meeting the ever-evolving needs of the academic community.

8. References

[1] Lawrence Webley, Tatenda Chipeperekwa & Hussein Suleman 2011 , *Creating a National Electronic Thesis and Dissertation Portal in South Africa University of South Africa*

[2] Jung-Ran Park 2009, *Metadata Quality in Digital Repositories: A Survey of the current State of Art*

[3] 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#*

[4] H Van de Sompel, ML Nelson, C Lagoze 2004 *Resource Harvesting for OAI-PMH Framework*

[5] Soumen Teli 2015, *Metadata harvesting from selected Digital Repositories in India: A Model to Build a Central Repository India*

[6] Edward A. Fox, Gail McMillan, Hussien Suleman, Macross A. Goncalves, Ming Luo 2017. *Networked Digital Library of Theses and Dissertations*

[7] *Information and Documentation- The Dublin Core Metadata element set. BS ISO, B Standard 2009*

[8] *The Dublin Core Metadata Element Set 2001. National Information Standards Organisation, Maryland*

[9] Virginia Braun, Victoria Clarke 2012 *Thematic Analysis*

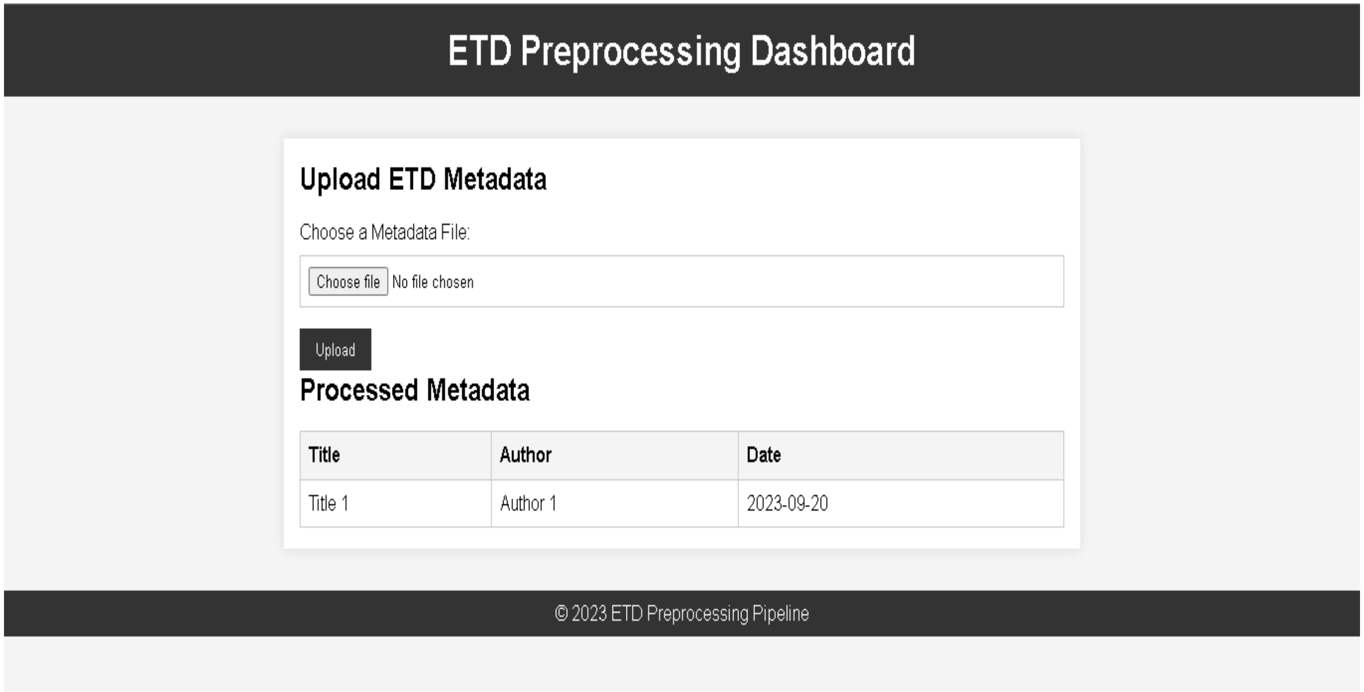
[10] Smith , Mckenzie 2003 *Dspace: An Open Source Dynamic Digital Repository, Volume 9 Number 1, Corporation for National Research initiatives.*

9. Appendices

Appendix A: Interview guide

INTERVIEW GUIDE

We are a team of Fourth Year Students at the University of Zambia carrying out a Project called The Zambia National ETD Preprocessing Pipeline which is a software tool aimed to improve the accessibility of ETDs generated by Zambian HEIs and will consistently format metadata and augment missing metadata using existing machine learning models as it has been noted number of Higher Education Institutions (HEIs) in Zambia offer advanced postgraduate programmes that automatically result in the publication of theses and dissertation manuscripts. The ETDs for a number of HEIs in Zambia are archived and, subsequently made available, via Institutional Repositories (IRs). Downstream services such as National ETD portals and global portals are generally used to aggregate metadata originating from such IRs, in order to provide end users with a unified interface for accessing ETDs from different sources.



ETD Preprocessing Dashboard

Upload ETD Metadata

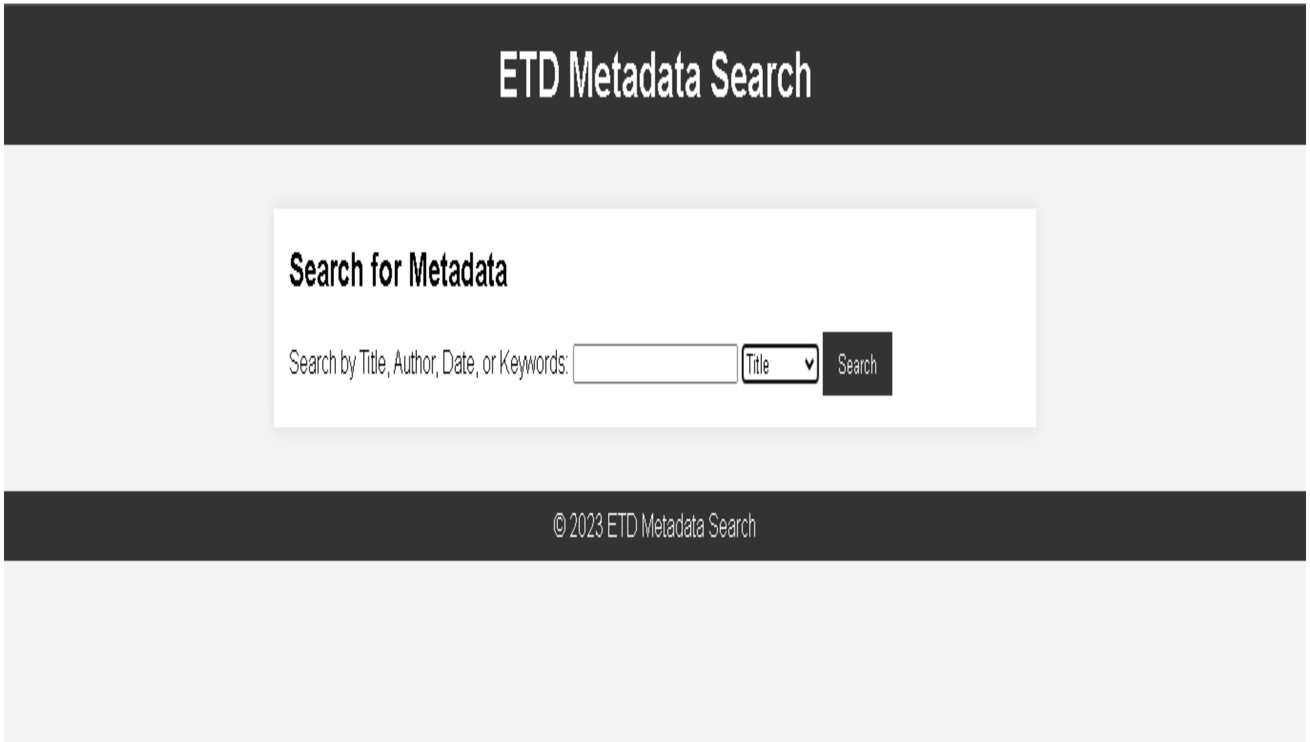
Choose a Metadata File:

No file chosen

Processed Metadata

Title	Author	Date
Title 1	Author 1	2023-09-20

© 2023 ETD Preprocessing Pipeline



ETD Metadata Search

Search for Metadata

Search by Title, Author, Date, or Keywords:

© 2023 ETD Metadata Search

‘

Background:

1. Briefly describe your role and responsibilities related to ETD Metadata and management?

Understanding the ETD Metadata Ingestion Process:

2. Can you describe the process of creating and managing ETDs at the University of Zambia and what metadata standards are followed in the process?

Identification of inconsistencies:

3. In your experience, what are some common inconsistencies or issues that you have encountered in ETD Metadata?
4. Are there specific metadata fields or elements that tend to have more inconsistencies than others?

Causes of inconsistencies:

5. What do you think are the main reasons behind these inconsistencies in ETD Metadata?
6. Are there any specific challenges that make it difficult to maintain consistent metadata ?

Impact of inconsistencies:

7. How do metadata inconsistencies affect the discoverability and accessibility of ETDs in your digital repository ?

Software Interface:

8. Provided with an interface like the one being shown, what changes would you recommend be made to it?
9. Is the interface user-friendly for metadata input? Yes or no
10. Would you appreciate the software tool having customizable metadata templates?
11. How would you want to have access to this metadata?

a) Would you want to access the metadata record as a whole document/theses or dissertation OR
b) Would you rather want to have access to the metadata in particular

12. What order would you prefer it to be in? E.g Author to title to Publisher
13. What casing would you recommend be used for the presentation and easy readability of the metadata?

Future improvements:

14. Looking ahead, what do you think can be done to further improve the consistency of ETD metadata across the academic community?
15. Are there any technologies that you believe could play a role in mitigating metadata inconsistencies?

Thank you for your time and contributions.