Rule-Based Automatic Generation of Electronic Theses and Dissertations Metadata

By

Frazer Nyambe (2018217071)

Nkole Mulenga (2018154818)

Richard Mufuzi (2018230069)

Geoffrey Ngoma (2018211536)

Supervisor:

Dr. Lighton Phiri

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

THE UNIVERSITY OF ZAMBIA

LUSAKA

2022

Abstract

In Universities postgraduates are required to submit theses and dissertations upon the completion of their studies. These are electronically uploaded to the repository and their metadata is manually generated and entered by the authorised people. The manual upload of theses and dissertations has led to the missing of metadata about the writings which makes it difficult for the lecturers and general public (students, postgraduates etc.) to access certain metadata elements from some records on the repository. Theses and dissertations are known to be the rich, unique source of information and hence they need to be paid attention to when uploading them. In an attempt to find a solution to this problem, this paper suggested and made use of automatic generation of metadata to identify the missing inputs about the writings in the ETDs section. Identification of the missing metadata elements was done by harvesting the metadata from the UNZA repository using the Open Archives Protocol for Metadata Harvesting (OAI-PMH), a widely adopted approach to allow harvesting of metadata. [8] This involved the pulling of harvested metadata using OAI-PMH URL validator, an implemented git bash script to download the records and carrying out a data analysis using an Excel Spreadsheet. Identifying the source of missing metadata elements from the manuscripts was achieved by reading through the Directorate of Research and Graduate Studies (DRGS) guidelines and then, randomly sampling out 60 Electronic Thesis and Dissertations (ETDs) from the UNZA repository. To determine the appropriate extraction method, the acknowledgements pages were first extracted from the 4149 PDF files, then converted to text and finally loaded to a pandas dataframe. Furthermore, rule-based matching techniques such as Spacy were used in a python script to extract the contributor (advisors) metadata details. Observably, excel analysis showed that only Eleven Dublin core elements were exported from openrefine out of a total of fifteen standard Dublin core elements. In addition to that, it was clearly observed that metadata elements such as contributor, source, coverage and rights were highly missing. After undertaking the analysis of the DRGS guidelines and the randomly sampling of 60 records from 12 schools, the major outcomes of possible elements drawn from the analysis showed that the metadata elements are found in the Approval and Acknowledgement section of the manuscript, but mostly on the Acknowledgements. It was observed that while trying to extract the supervisor details from the acknowledgements, the software library leaves out the salutation for the names. This is because SpaCy has a pre-set figure of speech that is capable of identifying the name from the sentences. In addition to that, records that never had the supervisor details were automatically skipped by the script. In Conclusion, the automatic extraction of the metadata from the manuscript is more effective as compared to the manual process. This conclusion was drawn based on the evaluation tested using the natural language processing metrics such as BLEU scores which take in the weight based on the human generated results versus the machine generated results.

Acknowledgements

Firstly we would love to thank God for granting us good health, knowledge, wisdom and making it possible for us to work together throughout this project as a team. We wish to thank our supervisor Dr. Lighton Phiri for being readily available for this group during the entire period of this Research project. His positive suggestions and reshaping of this research project has resulted in significant improvements.

We also wish to extend our gratitude to our fellow ICT 4014 students for the help rendered towards this project especially during presentation rehearsals and coding or script related milestones. Thank you all members of staff in the University Of Zambia, School Of Education, Department of Library and Information Science for the effort put in, the interactions and corrections shared to us at the time of presentations.

Many thanks to individual group members' effort put in towards the completion of each and every sub activities building up to this research project.

Lastly, we would like to thank our families for their support. Their belief in us has kept our spirits and motivation high during this process.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
AGETDM	AUTOMATIC GENERATION FOR ELECTRONIC THESES AND DISSERTATIONS METADATA
CRISP	CROSS INDUSTRY STANDARD PROCESS
CSV	COMMA - SEPARATED VALUES
DM	DATA MINING
ETDs	ELECTRONIC THESES AND DISSERTATIONS
HEIs	HIGHER EDUCATION INSTITUTIONS
IRs	INSTITUTIONAL REPOSITORIES
MS	METADATA STANDARD
NETDP	NATIONAL ETDs PORTAL
NLP	NATURAL LANGUAGE PROCESSING
OAI-ORE	OPEN ARCHIVE INITIATIVE OBJECT REUSE AND EXCHANGE
OAI-PMH	OPEN ARCHIVE INITIATIVE PROTOCOL FOR METADATA HARVESTING
XML	EXTENSIVE MARKUP LANGUAGE
DCMI	DUBLIN CORE METADATA INITIATIVE
AI	ARTIFICIAL INTELLIGENCE
NLTK	NATURAL LANGUAGE TOOLKIT
NER	NAMED ENTITY RECOGNITION

CHAPTER 1

1. Introduction

Since the year 2010, the University of Zambia has been ingesting ETDs into the repository for the purpose of research and as an institutional custom. Initially the association of metadata elements has been inconsistent due to some human factors. This project is about the automatic generation of metadata elements from the uploaded documents or records on the repository. Therefore, this research project has a solution to the problem which is based on prior work aimed at the implementation of rule-based techniques for classifying ETDs, and is about implementing appropriate techniques to automatically generate the missing metadata elements from the pdf manuscripts using tried and tested libraries and/or frameworks such as SpaCy and Grobid. The primary focus is to improve the quality of metadata harvested from the UNZA repository by comparatively evaluating the effectiveness of the solution using natural language processing metrics.

1.2. Background of Study

Research has been conducted pertaining to the Automatic Classification of Digital Objects for Improved Metadata Quality of Electronic Theses and Dissertations in Institutional Repositories (IRs). Furthermore, Improved Discoverability of Digital Objects in Institutional Repositories using Controlled Vocabularies also preceded [1]. This project is based on prior work conducted, thus it is important to note that the implementation of a prototype Zambian National ETD portal is meant to archive ETD generated by HEIs in Zambia [8]. In order to ensure that quality metadata harvested from HEI IRs by National ETD portals is comprehensive, it becomes necessary to automatically generate missing metadata.

1.3. Problem Statement

A number of Higher Education Institutions (HEIs) in Zambia offer advanced postgraduate programmes that automatically result in the publication of theses and dissertation manuscripts. The quality of ETDs metadata elements is poor and prior work [8] [1] has highlighted it in the findings. Poor quality of metadata has broad implications for downstream services that automatically harvest ETDs. In particular, downstream services such as national ETD portals and global portals tend to be adversely affected by low quality and incomplete ETD metadata associated with the uploaded manuscripts. Hence, content discovery in such portals becomes a problem to people who try to access ETDs.

A potential solution to poor quality and incomplete metadata elements associated with the UNZA repository that was highlighted by prior work. The solution to the problems involves identifying appropriate techniques to automatically generate the metadata using tried and tested libraries and/or frameworks such as Spacy and Gensim. Hence the focus of this research is on the missing metadata elements. Figure 1.1 shows an example record's screenshot with full metadata elements and from figure

1.1, we could clearly see that some of the metadata elements are missing, like for instance the supervisor details, and department details.

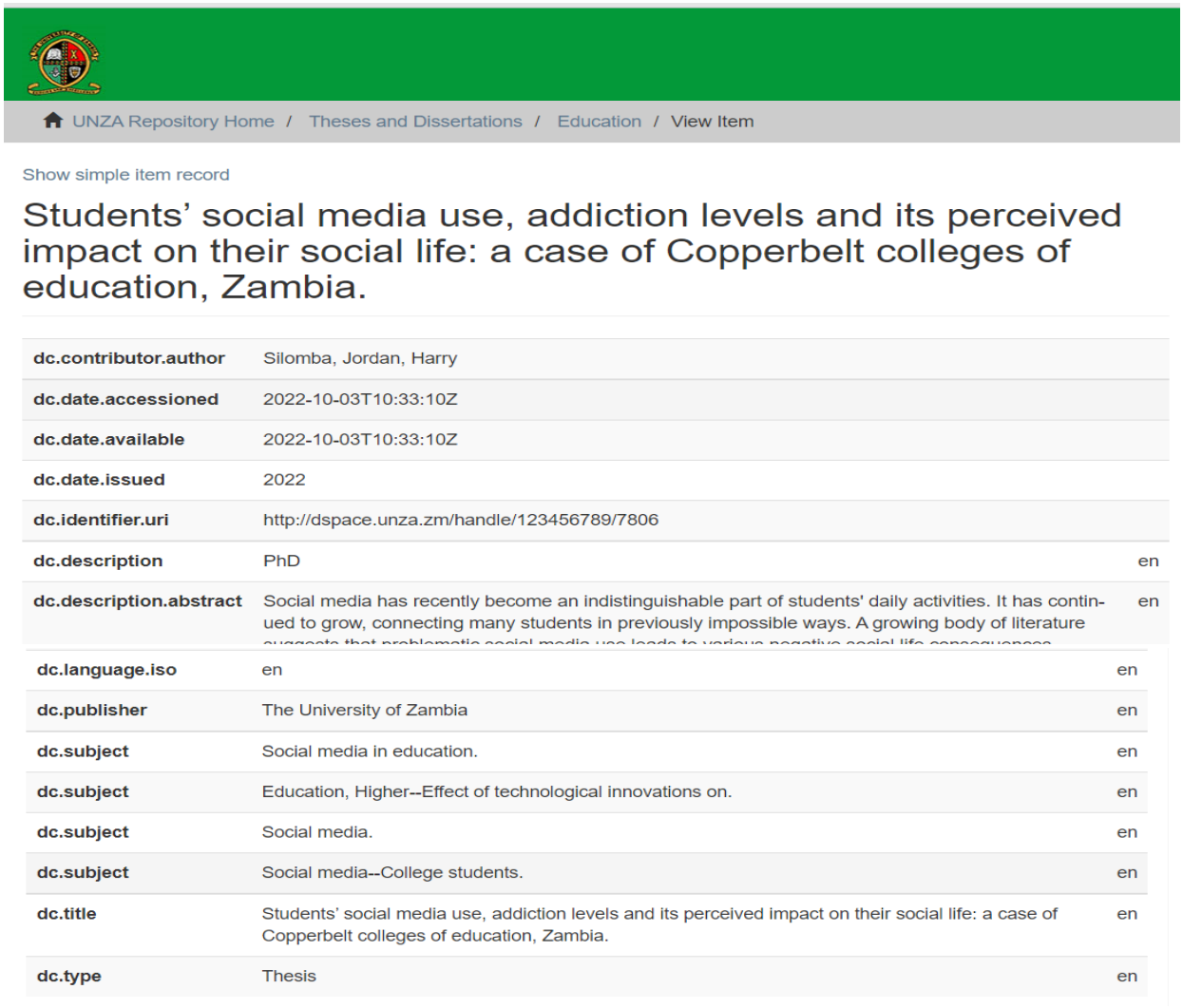


Figure 1.1.0: Screenshot of the UNZA Repository record with full metadata elements

1.4. Objectives of Study

1.4.1 Main Objective

The main objective of this project is to improve the quality of metadata harvested from external repositories by automatically generating missing metadata elements.

1.4.2 Specific Objectives

1. To identify common missing ETD metadata elements.

2. To identify the source of missing ETD metadata elements.

3. To determine the appropriate extraction method for missing ETD metadata from PDF manuscripts.

1.5 Research Questions

1. What common missing ETD metadata elements are to be identified in the PDF manuscripts?

2. Which part of the ETD PDF manuscripts contains the metadata?

3. How is the design and implementation of software tools useful in the extraction of metadata from PDF manuscript?

1.6 Ethical Considerations

Ethical considerations in any research have been identified or specified as one of the most important parts of the research. Researchers have identified research ethics as the core aspect of research and the foundation of a research design, an integral part of the research that needs to remain at the forefront of the research work. The importance therefore of the research ethics cannot be understated. The University of Zambia (UNZA) like any other institution has an Institutional Review Board (IRB) that has properly structured set of research ethical guidelines that every researcher conducting research at the university must follow, the IRB is responsible for ensuring the safety of the human participants and prevents the violation of their human rights. The IRB at the university reviews the methodologies and aims of the research study and ensures that the ethical guidelines are followed and if the conditions of the IRB are not fully met the research study will have to be amended.
This project intends to shield users' information from any actions that may be perceived as a threat to their identity and privacy. Some of the ethical issues related to this project may include the following:

1.6.1 Informed consent

With the use of natural language processing, this research project will work in line with the terms and conditions for the particular software tools that will be used for design and implementation. This will consist of using free and open source software.

1.6.2 Respect of privacy

Privacy breaches disturb trust and run the risk of diluting or losing security. It is a show of disrespect to the law and a violation of ethical principles. Therefore, this project will ensure that only people who ask for copyright permission will be able to share or publish the work.

CHAPTER 2

2. Related Work

2.1. Identification of common missing metadata elements for improved metadata quality

2.1.1. Author name disambiguation

For any work of literature, a fundamental issue is to identify the individual(s) who wrote it, the supervisor(s), institution name(s), and year of publication and conversely, to identify all of the works that belong to a given individual [12]. Author name disambiguation comprises four distinct challenges:

Under distinct one, a single individual may publish under multiple names this includes (a) Orthographic and spelling variants, (b) Spelling errors, (c) Name changes over time, as may occur with marriage, religious conversion, or gender reassignment, and (d) The use of pen names. In distinct two, many different individuals have the same name—in fact, common names may comprise several thousand individuals. Distinct three, the necessary metadata is often incomplete or lacking entirely. For example, some publishers and bibliographic databases did not record authors' first names, their geographical locations, or identifying information such as their degrees or positions. In the final distinct, an increasing percentage of scholarly articles are not only multi-authored, but represent multi-disciplinary and multi-institutional efforts.

Nevertheless, our research will focused on the standardisation of extracting the first name and last name for the supervisor details identified by the software libraries' Named Entity Recognition like spacy, Natural Language Toolkit (NLTK), Stanford NER[6] etc. , unlike just focusing on the author name ambiguity.
SpaCy is known for its industrial-strength natural language processing library in Python. It has been written in Cython which is a superset of Python programming language with C-like performance.

This is similar to our work in the use of different name variations for author names. Unlike the cited scholarly research, our research project mainly focused on identifying the supervisor details despite the name ambiguity of the authors.

Stanford NER is a Java implementation of a Named Entity Recognizer. Stanford NER is also known as CRFClassifier. The software provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models. That is, by training your own models on labelled data, you can actually use this code to build sequence models for NER or any other task.
NLTK (Natural Language Toolkit) is a great Python package that provides a set of natural languages corpora and APIs of wide varieties of NLP algorithms. NLTK comes along with the efficient Stanford NER implementation.

2.1.2. Automatic classification of Digital Objects for improved Quality of Electronic Theses and Dissertations in Institutional Repositories

A key feature of IRs is facilitating effective discoverability of content through search and browsing functionalities. In order for IRs to cluster related digital objects together, controlled vocabulary sets are used [7]. The vocabulary sets provide a mechanism for presenting a restricted set of terms during the ingestion of digital objects into IRs. Digital objects indexed in the ACM Digital Library4 are normally tagged using the ACM Computer Classification System (CCS). Some PubMed Central-5 articles are tagged with the Medical Subject Heading (MeSH) classes. Existing literature highlights effective discoverability and improved interoperability as being the key advantages of using controlled vocabulary.

The Library of Congress Subject Headings (LCSH) terminology is currently used by UNZA [1]. However, ETDs that are kept in IR are written by graduate students from various academic fields. UNZA is made up of 13 faculties, each of which has a variety of departments that are linked to restricted terminology that is exclusive to that field. For instance, the Biological Sciences, Chemistry, Computer Science, Mathematics, Physics, and Geography and Environmental Studies departments make up the School of Natural Sciences. Each of these divisions may be connected to a wide range of controlled terminology.

In our research, the focus was on using software libraries that would automatically identify the predefined named entities based on the figure of speech from the document sentences and coming up with predefined matching phrases, unlike just the arrangement of words and phrases. Work of spaCy Named Entity Recognition as already eluded in the previous literature 2.1.2.

This related literature is similar to our work due to its use of controlled vocabularies to identify the subject headings for the research work and the improvement of metadata quality, but differ in the use of supervised machine learning in the extraction of elements from literature.

2.1.3 Dublin Core Elements

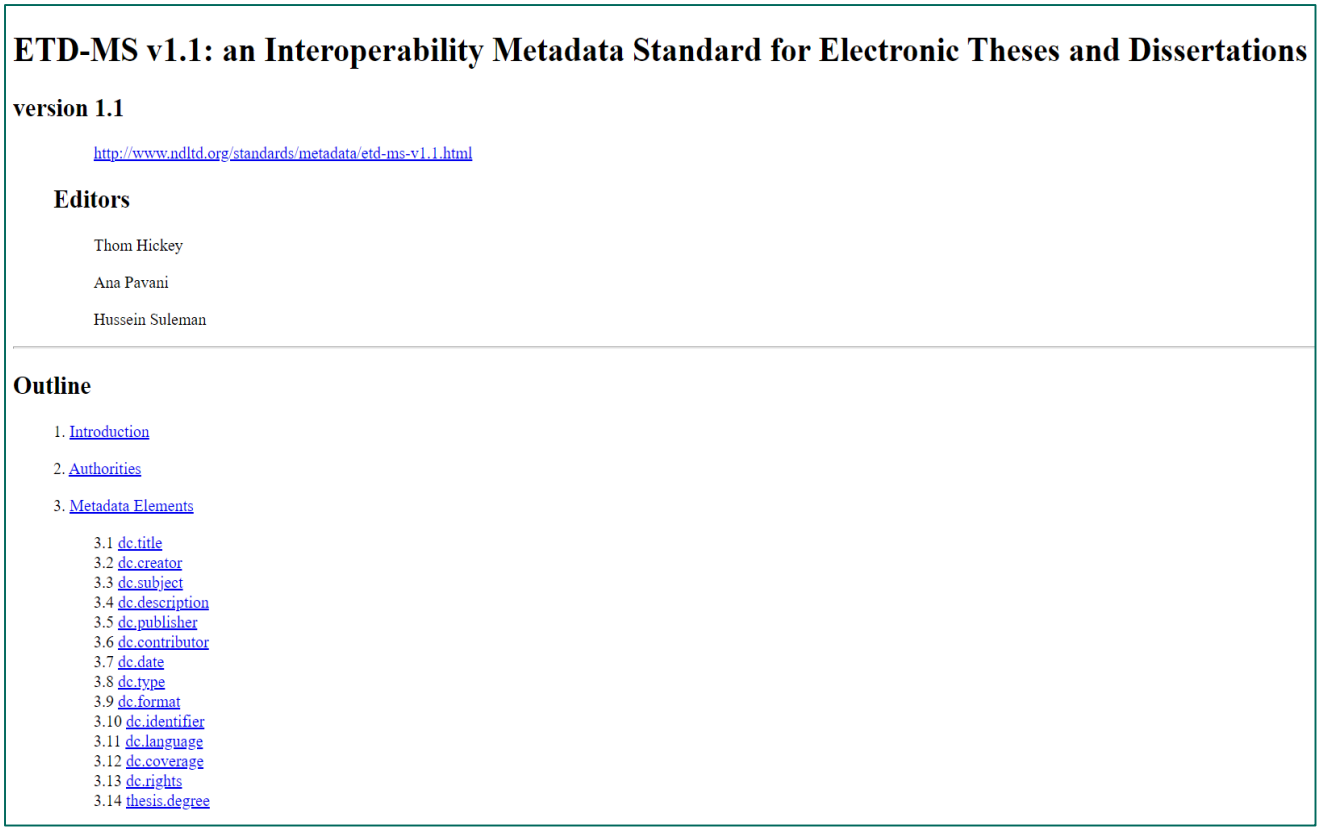
The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. The name "Dublin" is due to its origin at a 1995 invitational workshop in Dublin, Ohio; "core" because its elements are broad and generic, usable for describing a wide range of resources [2]. The fifteen element "Dublin Core™" described in this standard is part of a larger set of metadata [17] vocabularies and technical specifications maintained by the Dublin Core™ Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms [DCMI-TERMS], also includes sets of resource classes (including the DCMI Type Vocabulary [DCMI-TYPE]), vocabulary encoding schemes, and syntax encoding schemes. The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCMI Abstract Model. The fifteen Dublin Core elements are defined in the table below. Table 2 shows the fifteen Dublin Core elements with their definitions.

Table 1. The 15 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

2.1.4 ETD-MS metadata

ETD-MS defines a standard set of metadata elements used to describe an electronic thesis or dissertation. Institutions dealing with electronic theses and dissertations have all developed their own standards or adapted existing metadata standards [3]. In line with the documentation, these metadata standards all attempt to describe the author, the work, and the context in which the work was produced in a way that will be useful to the researcher as well as the librarians and/or technical staff maintaining the work in its electronic form. However, the source document is not a replacement for the metadata schemes developed for a particular university or environment. Rather, a document that should be used as a guideline to develop a faithful crosswalk between local metadata standards and a single standard used for sharing information about ETDs. Figure 2.1.0 and figure 2.2.0 showing a screenshot for ETD-MS metadata standard with the prescribed metadata elements.



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 2.1.0: Screenshot for ETD-MS metadata standard

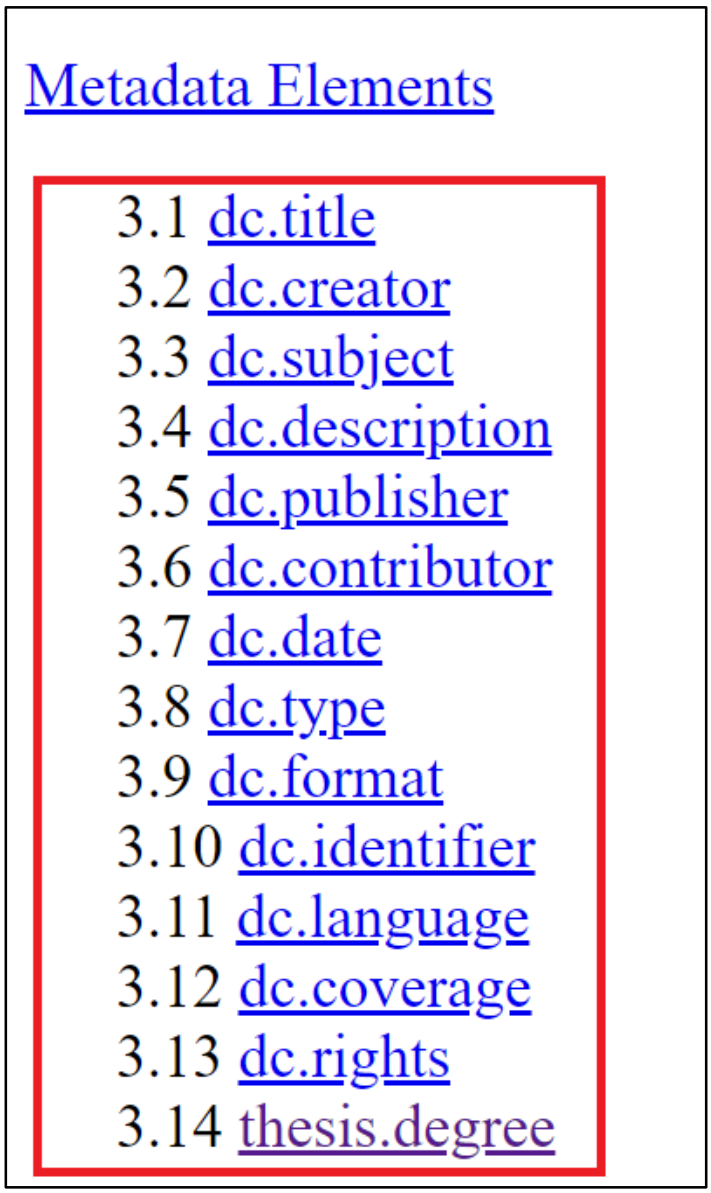


Figure 2.2.0: ETD-MS metadata elements

2.2. Identification of the source of missing ETD metadata elements

2.2.1. Improved Discoverability of Digital Objects in Institutional Repositories Using Controlled Vocabularies

Identification of the source of the missing ETD metadata can be done by controlled vocabulary and authority control, which are popular techniques that are used to enhance access to bibliographic materials. Controlled vocabularies are well-organised words and phrases that are used to index digital content and subsequently facilitate retrieval of the content through searching and browsing. Subject headings are a form of controlled vocabulary that are used to describe topics associated with digital content, making it possible for content related content to be grouped together[1]. While generic subject headings such as the Library of Congress Subject Headings are widely used, there are other domain-

specific subject headings, popular with academic databases. For instance, the Medical Subject Headings terms are used in the medical field, and the ACM Computing Classification System ontology is commonly used in computing disciplines. Prior work on the identification of the source of ETD metadata elements focused on empirically determining the implications of sparing use of subject headings and, additionally, identifying potential domain-specific subject headings that can be incorporated into IRs. Furthermore, this demonstrated the positive effect subject headings have on the overall usability of IRs.

Similarly to this research work, the focus of our research was also on scanning the whole document for common missing metadata elements using controlled vocabularies sets from the manuscripts using cross match of UNZA repository and interoperability ETD-MS metadata elements. However, the Directory of Research and Graduate Studies guidelines specifically the preliminary section was used as a basis to identify the areas of extracting the metadata elements from the manuscripts. Below are some relatable literature information:

Reading order resolving

A PDF file contains by design a stream of strings that undergoes extraction and segmentation process. As a result, we obtain pages containing characters grouped into zones, lines and words, all of which have a form of unsorted bag of items. The aim of setting the reading order is to determine the right sequence in which all the structure elements should be read. This information is used in zone classifiers and also allows to extract the full text of the document in the right order. An example document page with a reading order of the zones is shown in Figure below.

Content classification

The goal of content classification is to determine the role played by every zone in the document. This is done in two steps: initial zone classification (A4) and metadata zone classification (B1).

The goal of initial classification is to label each zone with one of four general classes: *metadata* (document's metadata, e.g. title, authors, abstract, keywords, and so on), *references* (the bibliography section), *body* (publication's text, sections, section titles, equations, figures and tables, captions) or *other* (acknowledgments, conflicts of interests statements, page numbers, etc.).

The goal of metadata zone classification is to classify all *metadata* zones into specific metadata classes: *title* (the title of the document), *author* (the names of the authors), *affiliation* (authors' affiliations), *editor* (the names of the editors), *correspondence* (addresses and emails), *type* (the type specified in the document, such as "research article", "editorial" or "case study", *abstract* (document's abstract), *keywords* (keywords listed in the document), *bib info* (for zones containing bibliographic information, such as journal name, volume, issue, DOI, etc.), *dates* (the dates related to the process of publishing the article).

2.2.2. Automatic extraction of structured metadata from scientific literature

Academic literature is a very important communication channel in the scientific world as it helps different people to reference the available work from the repositories. This information is extracted from different sources. In this section, a few sources are stated to give a view of what automatic extraction of structured metadata is about.

Cermine

Cermine is a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form [14]. The extraction algorithm proposed by CERMINE performs a thorough analysis of the input scientific publication in PDF format and extracts:

A rich set of document's metadata, a list of bibliographic references along with their metadata, and a structured full text with sections and subsections (currently in experimental phase).

This literature uses structured data that is also used by our research. The only difference is in the extraction techniques, this related work used machine learning techniques and we used rule based techniques.

University libraries often provide public access to digital libraries of ETDs. Similarly, some universities scan or digitise older theses and dissertations in order to provide electronic access to these works. To the best of our knowledge, scan bank was the first manually annotated dataset for figures and table extraction for scanned ETDs.

2.3. Determination of the appropriate extraction method for missing ETD metadata from pdf manuscripts

2.3.1. A Method of Extracting Metadata Information in Digital Books

Extraction of metadata focuses on the three main formats, mainly: TXT, PDF, and HTML [10]. All of which have their own characteristics in terms of standards, size, and analytical methods. Algorithms analyse their features, obtaining basic metadata information. The algorithm that extracts TXT metadata uses a method based on supporting vector machine models. The HTML parser is open source, and HTML is a tree structure. Text content is stored in various labels. The document's text is extracted from HTML, and its format is not much different from TXT. Therefore, we can use the method of supporting vector machine models to extract HTML metadata. The method of extracting metadata from PDF documents is divided into two steps. Firstly, the PDF document is converted to text; secondly, the method based on supporting vector machines is used to extract the metadata from the converted text data. The text format can be obtained after filtering with open source libraries, so that the open source database can be used to build a model of support vector machine simulation for extracting the metadata of the document.

There are a range of tools available for detecting and extracting specific page numbers from documents. They are mostly library packages in python to easily use the tools for extraction. They take the PDF as

input and detect the specific pages across the PDF. One can also give the page number as input to detect specific metadata attributes inside the PDF. Some of the tools that we have studied for the purpose of use in our project are PDFTK and PyPDF2.

PyPDF2 is an open-source package in Python which is used for performing major tasks on PDF files such as extracting the document-specific information in textual format, merging the PDF files, splitting the pages of a PDF file, adding watermarks to a file, encrypting and decrypting the PDF files, etc. It does not recognize the layout of tables and it just extracts the data in text format [9].This package is useful for extracting text from many PDFs which can later be used in Natural Language Processing applications. The only difference between our research and this related is that, the related extracts the html content as well unlike just the focus on PDF and TXT formats.

PDFTK which is another open-source package in Linux used to extract certain pages from one or more PDF files into a new PDF [13]. It is a free tool for processing PDF files, including their splitting and merging, decryption and encryption, and bursting into single pages.

2.3.2. Improved Discoverability of Digital Objects in Institutional Repositories Using Controlled Vocabularies

Interoperability is a computer system's ability to be interfaced with other external system services through the standardised use of predefined data formats and communication protocols [1]. Suleman states that in the context of DLs, interoperability promotes openness, a key philosophy mandated by the Open Access movement. The use of international standards ensures that other tools and services are easily able to use and integrate the data. Prior work focused on the use of DLs that employ communication protocols to provide auxiliary services for facilitating the core functionalities associated with repositories—ingestion, management, search, and browsing of digital objects.

While this research will only make use of protocols for digital harvesting. For instance, communication protocols such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Open Archives Initiative Protocol for Object Reuse and Exchange (ORE) that are used for harvesting digital object metadata and bit streams, respectively [17]. This differs from our work because of its Integrating IRs with subject controlled vocabularies, but they both focus on scanning the whole document for common missing metadata elements using controlled vocabularies sets.

CHAPTER 3

3. Methodology

This section outlines research methods that were used in the study. It provides information about modern technologies for automatic generation of missing metadata from PDF manuscripts like software libraries that are associated with natural language processing techniques. The instruments that were used to collect relevant information needed for the design and implementation for the purpose of a comprehensive and broader approach are also described and the process or phases involved in conducting this study are included.

3.1. Identification of missing metadata elements

To identify the missing metadata elements we harvested the metadata from the UNZA repository using the OAI-PMH protocol. The pulling of harvested metadata was done using two methods, firstly; copying of the URL link for the theses and dissertations OAI-PMH data provider and pasting the link into OAI-PMH URL validator to download the 17 XML files. Secondly, we implemented a git bash script to download the 4149 properly formatted records excluding the 200 malformed records using resumption tokens composed of 100 records per file.

Furthermore, the downloaded 4149 records were uploaded to openrefine to orderly group the records and export the record files as Excel Spreadsheet for data analysis.

Data analysis was done on an Excel Spreadsheet to identify the associated metadata elements to the UNZA repository and a cross match was carried out between the UNZA repository and the ETD-MS metadata standard. In order to achieve the expected high quality metadata elements, an exemplar record was picked from the OpenUCT repository and cross matched with ETD-MS.

3.2. Identification of the source of missing ETD metadata elements

To identify the source of missing metadata elements from the manuscripts, two steps were taken, the first being the downloading of post graduate guidelines from the directorate of research studies (DRGS). We then read through the document, specifically the preliminary section which guides postgraduate students on how to format their theses and dissertation.

In the second step, we sampled 60 theses and dissertations from the UNZA repository, five from each school. The individual documents were then analysed to enable us to have possible metadata elements that we could work with.

Detailed analysis of results and steps drawn with snapshots taken from any of the sampled documents' wanted section(s) is mentioned in the Results and Discussions section below.

3.3. Determination of appropriate extraction method of missing ETD elements from the PDF manuscripts.

To determine the method of extraction for the source of truth for the missing metadata elements gathered at study #2, we had to extract the acknowledgements pages from the 4149 PDF files and convert them to text files first. Next we loaded the converted text files into a pandas dataframe. Furthermore, to identify the supervisor details from the text files, SpaCy library's named entities were used. Using the Rule-Based Approach for basic string matching, we came up with a python script to extract the identified supervisor details. To have a clear overview of the extracted supervisor details for each record, results were printed out with its specific aligned unique identifier.

Rule-based approaches make use of software libraries such as Gensim[5] and Spacy[16] to design a rule that extracts a text such as the supervisor details (contributor) e.g. Dr. Lighton Phiri.

However, to achieve this objective for study #3, various steps were taken that included creation of a dataset by downloading all the 4149 PDF documents from the repository using a bash script, then the aforementioned extraction methods above follows.

Detailed analysis, steps and programming processes are documented in the Results and Discussions section.

CHAPTER 4

4. System Design and Implementation

According to Creswell [18] research design involves plans, structures and strategies for conceiving investigation so as to obtain answers to research questions and control variance. In this study, the Cross Industry Standard Process for Data Mining (CRISP-DM) model (which is a process model that serves as the basis for a data science's focus of the project's methodological approach and how data mining takes place with all of the six phases) was utilised. However the study applied both the quantitative and qualitative research processes. Quantitative approach helped to quantify the problem by measuring data from the pdf manuscripts. These two research approaches on the other hand helped with data-interpretation and to study the behaviours as well as the defined variables. Figure 4.1.0 shows the main components of the design and implementation for this research project.

The six phases from the CRISP-DM model utilised for the purpose of Data collection, Data pre-processing and Rule-based model are as follows;

4.1 Business Understanding

With the clear understanding of the prior work that was conducted in the related work section, this research aimed at understanding the manual upload of ETDs into the UNZA repository so that we may have an idea on how the missing metadata elements from the PDF manuscripts can be identified and generated automatically.

4.2 Data Understanding

This phase involved steps in data collection as a way of giving answers to the research questions.

Step 1

A cross-match between the UNZA repository and an Open access institutional repository of the University of Cape Town's (OPEN UCT) metadata elements and a document analysis on the Electronic Thesis and Dissertations - Metadata Standard (ETD-MS) for the prescribed metadata elements.

Step 2

An analysis of a document for the DRGS postgraduates' regulations was carried out by reading through its sections; this was followed by randomly sampling 60 records from the different schools that have uploaded Electronic thesis and dissertations (ETDs).

Step 3

A rule-based technique specifically spaCy and other python library packages such PyPDF2, glob, pandas, and a bash script were used as extraction methods to automatically generate metadata from PDF files. These software tools were very useful and vital to this research in such a way that generation of missing inputs in the ETDs was easy and paved an easy way to use text generation metrics for evaluating the system.

4.3 Data Preparation

Just like the phrase "garbage in garbage out", Openrefine was employed on raw data for all common text pre-processing techniques: removal of stop words, punctuation marks, and numbers, stemming,

and handling of null values. The reason behind the use of this online tool is in its ability to group or merge multiple rows of a record into a single row, filtering or faceting record information as well as renaming of the record columns. The cleaned data was then exported to a Google spreadsheet for another data processing.

4.4 Modelling

The Rule-based model such as spaCy was key towards the extraction of the needed metadata elements e.g. contributor or advisor's details. Figure 4.1.0 (tabular illustration) below shows how metadata elements from ETDs get to be extracted, sourced and finally extracted by means of the model employed.

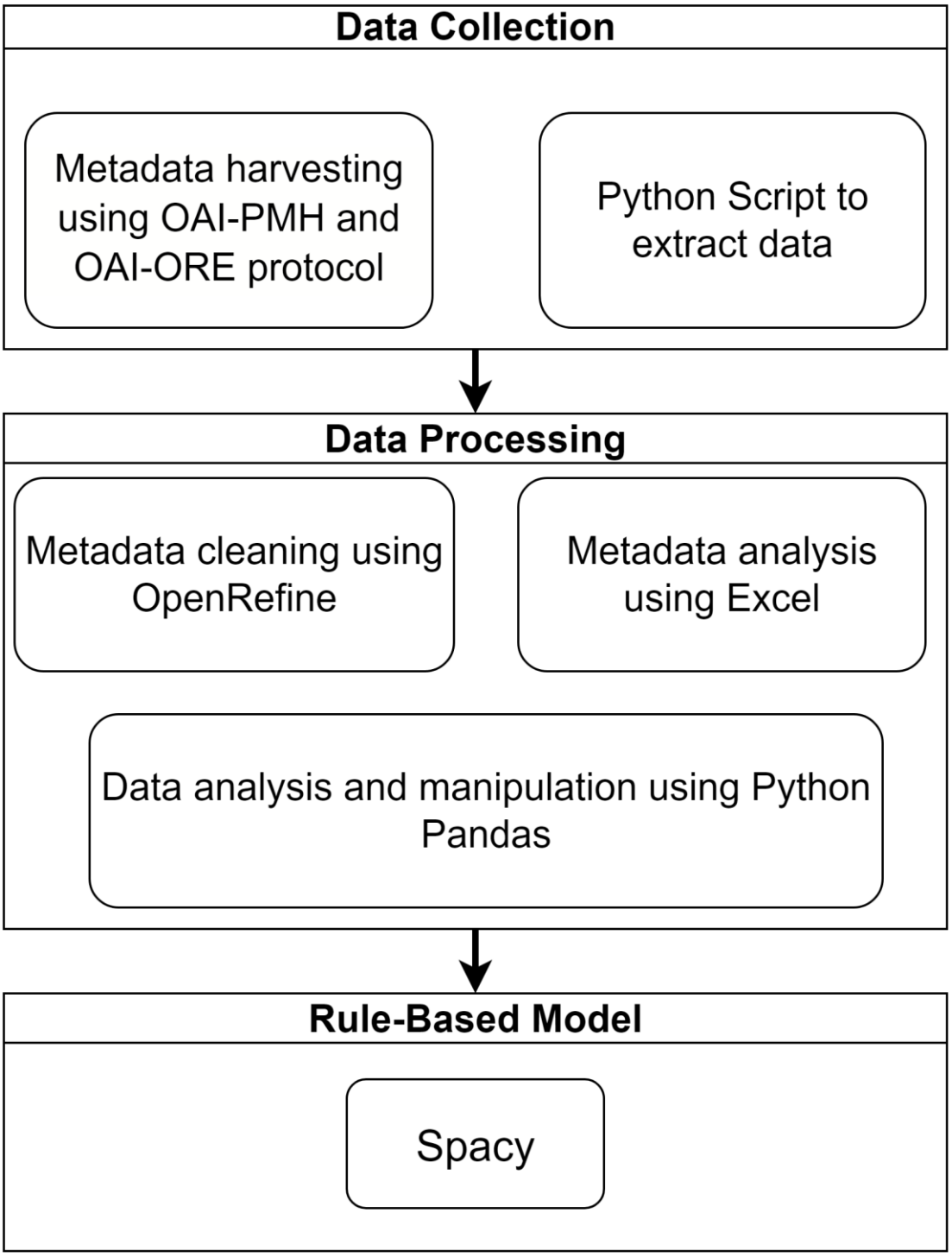


Figure 4.1.0: Design and implementation components

4.5 Evaluation

The Natural Language Processing text generation metrics to be specific Bleu scores and Rouge were used as means for this research work's evaluation. This was to assess their relative effectiveness by measuring the accuracy of the standard automatic generated estimators. In addition, the effectiveness

of feature combinations (generated solutions and reference solutions) were assessed to determine if the end goal is achieved.

4.6 Deployment

As an overview of future developments, developing the model APIs would really enhance the communication between the model and other systems.

CHAPTER 5

5. Results and Discussions

5.1. Identification of missing metadata elements

The University of Zambia has an online research repository that is used for the upload of scholarly research projects by postgraduate students. The repository is mainly updated with theses and dissertations done by postgraduate students upon completion of their scholarly research projects. Currently the repository has a total number of 4477 deposited records prone to change. To identify the missing metadata elements from the records various steps were carried out namely data harvesting, data extraction and data analysis.

5.1.1. Data harvesting

The harvesting of metadata from the records on the repository was done using the data provider. These records were harvested in a batch of hundreds using the resumption token as shown below in Figure 5.1.1 and Figure 5.1.2.

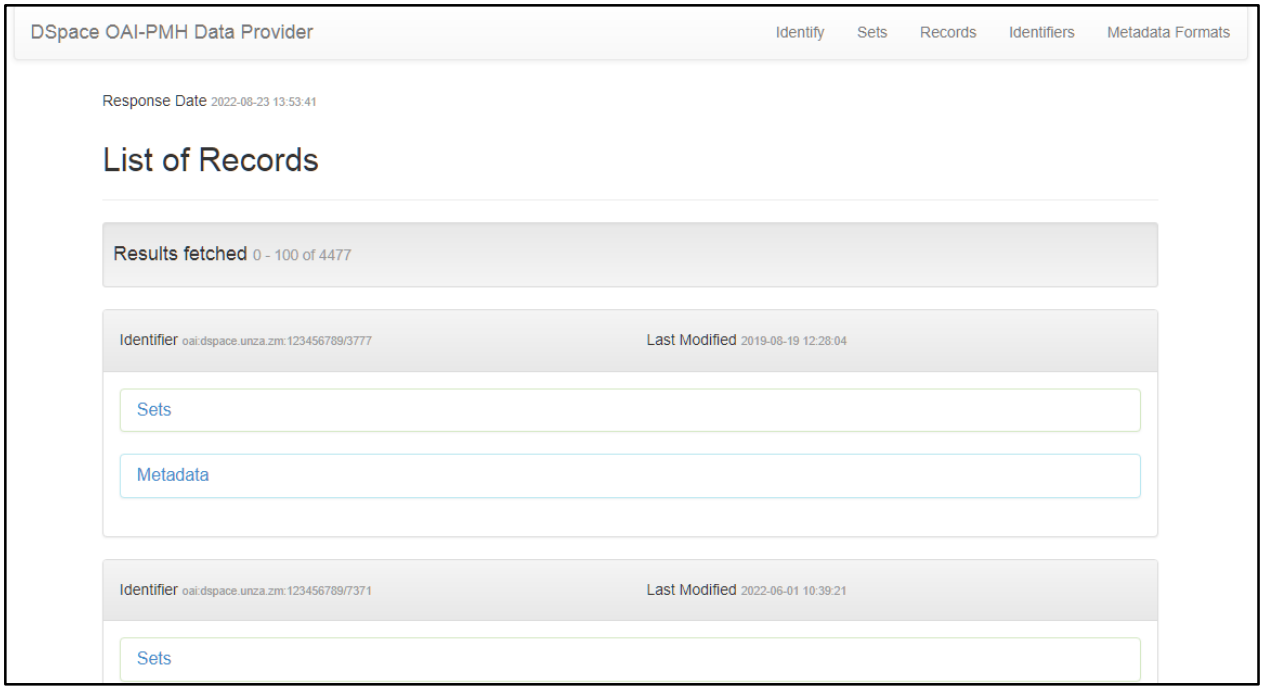


Figure 5.1.1: Records from DSpace

Figure 5.1.2: Metadata section from DSpace

5.1.2. Data Extraction

To extract the metadata elements we came up with a git bash script that was able to pull the metadata files from the repository. This script was using a "wget command" that requires the URL to the repository for each resumption token to download or pull the xml records. The downloaded record files are stored on the computer for further analysis. After the downloading and storing of the files on the computer, Openrefine was used to merge the xml files into one analysable file. Furthermore, Openrefine was used to merge each record's row into one, as well as renaming the record columns for the eleven Dublin core elements and lastly the merged records were exported to excel for better visualisation and analysis.

It was observed that Openrefine encountered errors when parsing the files from the two resumption tokens, namely for the records between 1500-1600 and 1600-1700. The encountered errors were due to malformation of records during the upload process. To deal with the errors, the resumption tokens with errors were skipped as shown in Figure 5.1.3 below. In addition to that, it was clearly observed that the records from the remaining 43 resumption tokens were properly formatted and easy to parse hence giving out the total number of 4149 records as shown in Figure 5.1.3. Figure 5.1.4, Figure 5.1.5, Figure 5.1.6, Figure 5.1.7 and Figure 5.1.8 show screenshots for the various steps that were taken in Openrefine.

	Script
2	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/100" | xmllint --format - > dspace_unza_zm-output_file-100.xml
3	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/200" | xmllint --format - > dspace_unza_zm-output_file-200.xml
4	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/300" | xmllint --format - > dspace_unza_zm-output_file-300.xml
5	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/400" | xmllint --format - > dspace_unza_zm-output_file-400.xml
6	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/500" | xmllint --format - > dspace_unza_zm-output_file-500.xml
7	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/600" | xmllint --format - > dspace_unza_zm-output_file-600.xml
8	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/700" | xmllint --format - > dspace_unza_zm-output_file-700.xml
9	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/800" | xmllint --format - > dspace_unza_zm-output_file-800.xml
10	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/900" | xmllint --format - > dspace_unza_zm-output_file-900.xml
11	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1000" | xmllint --format - > dspace_unza_zm-output_file-1000.xml
12	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1100" | xmllint --format - > dspace_unza_zm-output_file-1100.xml
13	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1200" | xmllint --format - > dspace_unza_zm-output_file-1200.xml
14	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1300" | xmllint --format - > dspace_unza_zm-output_file-1300.xml
15	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1400" | xmllint --format - > dspace_unza_zm-output_file-1400.xml
16	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1500" | xmllint --format - > dspace_unza_zm-output_file-1500.xml
17	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1600" | xmllint --format - > dspace_unza_zm-output_file-1600.xml
18	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1700" | xmllint --format - > dspace_unza_zm-output_file-1700.xml
19	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1800" | xmllint --format - > dspace_unza_zm-output_file-1800.xml
20	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/1900" | xmllint --format - > dspace_unza_zm-output_file-1900.xml
21	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/2000" | xmllint --format - > dspace_unza_zm-output_file-2000.xml
22	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/2100" | xmllint --format - > dspace_unza_zm-output_file-2100.xml
23	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/2200" | xmllint --format - > dspace_unza_zm-output_file-2200.xml
24	wget -O - "http://dspace.unza.zm/oai/request?verb=ListRecords&resumptionToken=oai_dc///com_123456789_18/2300" | xmllint --format - > dspace_unza_zm-output_file-2300.xml

Figure 5.1.3: Scripts for the 43 Resumption Tokens

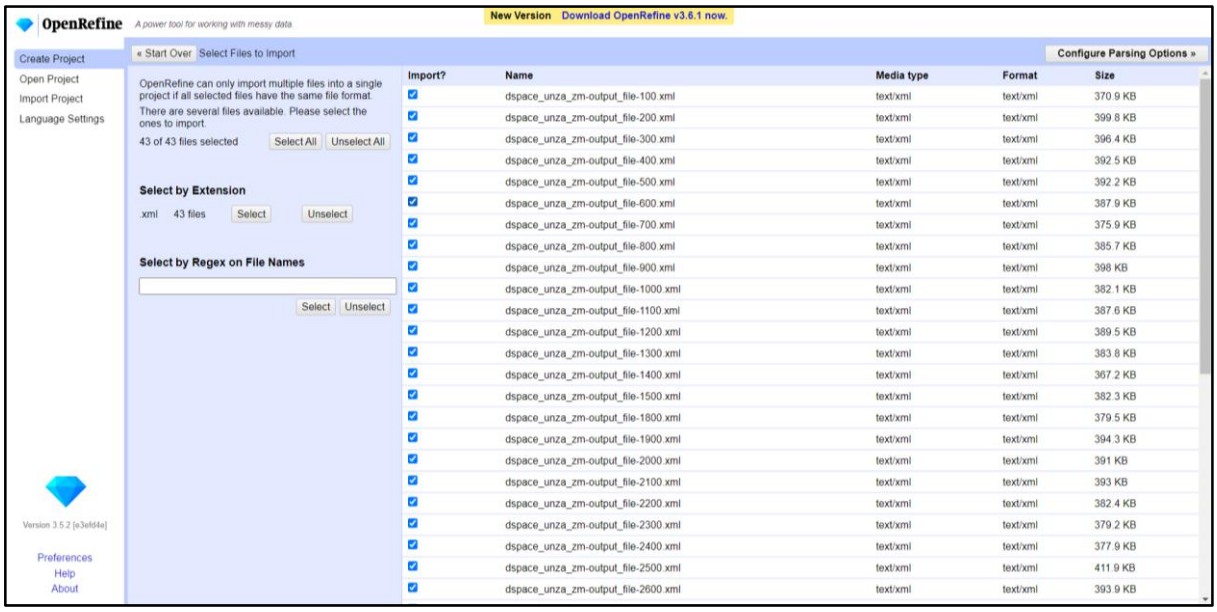


Figure 5.1.4: Imported 43 xml Files

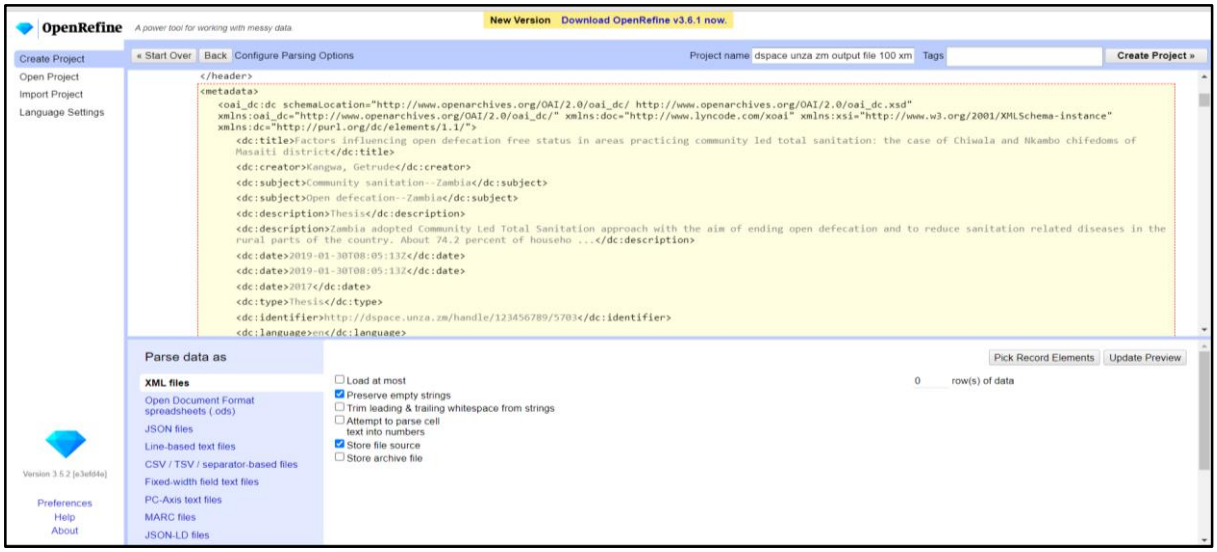


Figure 5.1.5: Specification of Parsing Data Paths

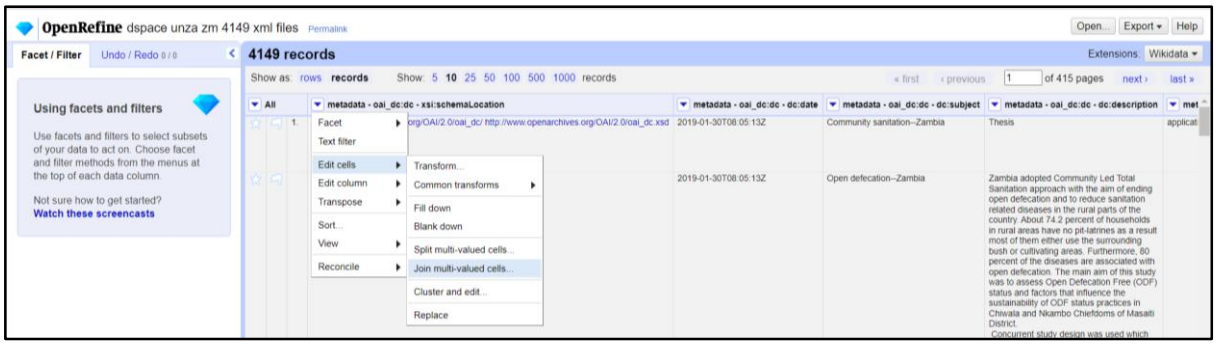


Figure 5.1.6: Joining Of Multiple Record Rows

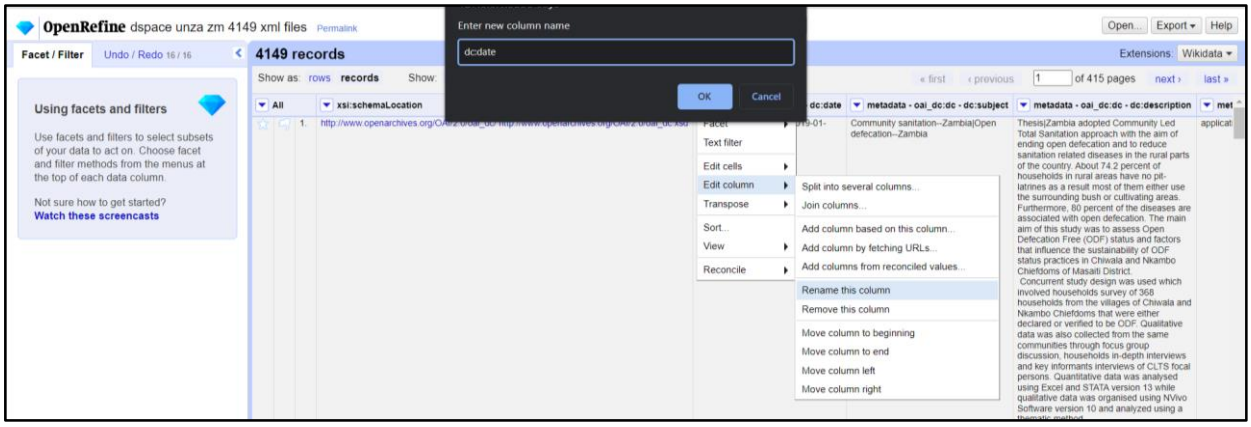


Figure 5.1.7: Renaming Of Record Columns

Figure 5.1.8: Export of the Joined Records to Excel

5.1.3. Data Analysis

Data was analysed in excel after being exported from Openrefine. This was done by manipulating it using a pivot table feature that enabled the grouping of the eleven Dublin core elements associated to the available records. The records were arranged in relation to the year of publication using a formula (trimming of the dc: date's ingestion date), these records were able to show the number of missing metadata elements that each of them consisted from a specific year.

Observably, only Eleven Dublin core elements were exported from openrefine out of a total of fifteen standard Dublin core elements that include; the publisher, format, creator, identifier, title, language, subject, relation, description, date, coverage, source, type, contributor and rights. The information given can be seen in Table 2 below. The zero values in red show the metadata elements that are associated to the records on the UNZA repository and the rest show the number of missing metadata elements.

Table 2. Analysis of missing Data in Excel Spreadsheet

dc: year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
PUBLISHER	28	636	736	396	171	332	8	5	5	7	10	3	1	**2338**
DESCRIPTION	3	14	23	86	13	6	0	0	0	1	2	0	0	**148**
RELATION	28	639	738	396	171	331	169	449	68	190	332	222	413	**4146**
SUBJECT	0	1	1	1	1	2	1	0	0	0	2	0	0	**9**
LANGUAGE	0	4	2	0	0	2	0	10	0	0	1	1	2	**22**
TYPE	0	4	4	1	0	5	0	3	0	3	2	0	0	**22**
TITLE	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
IDENTIFIER	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
CREATOR	0	1	0	0	0	3	0	0	0	0	1	0	0	**5**
FORMAT	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
DATE	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
CONTRIBUTOR	28	639	738	396	171	333	169	449	68	190	333	222	413	**4149**
SOURCE	28	639	738	396	171	333	169	449	68	190	333	222	413	**4149**
COVERAGE	28	639	738	396	171	333	169	449	68	190	333	222	413	**4149**
RIGHTS	28	639	738	396	171	333	169	449	68	190	333	222	413	**4149**

5.1.3.1. Results of varying associated metadata by year

Table 3 shows the distribution numbers of the records missing in each of the 15 Dublin core elements from 2010 to 2022 when the records were uploaded to the repository. It is easier to see how many records in a particular year were associated with the 15 Dublin core metadata elements and the ones that were missing. In addition to that, it was clearly observed that metadata elements such as contributor, source, coverage and rights were totally missing.

From the graphical line representation of the results in Figure 5.2.0, there has been a variation in the association of metadata elements from the initial year of ingestion (2010) to the current year (2022). It was observed that between 2010 and 2022 some metadata elements like contributor, source, coverage and rights were missing from all the records. On the other hand, the remaining elements were identified but were also missing from the other years. Figure 5.1.9 shows the number of missing metadata elements in percentages per year and Figure 5.2.0 also shows the graphical representation of the missing metadata elements.

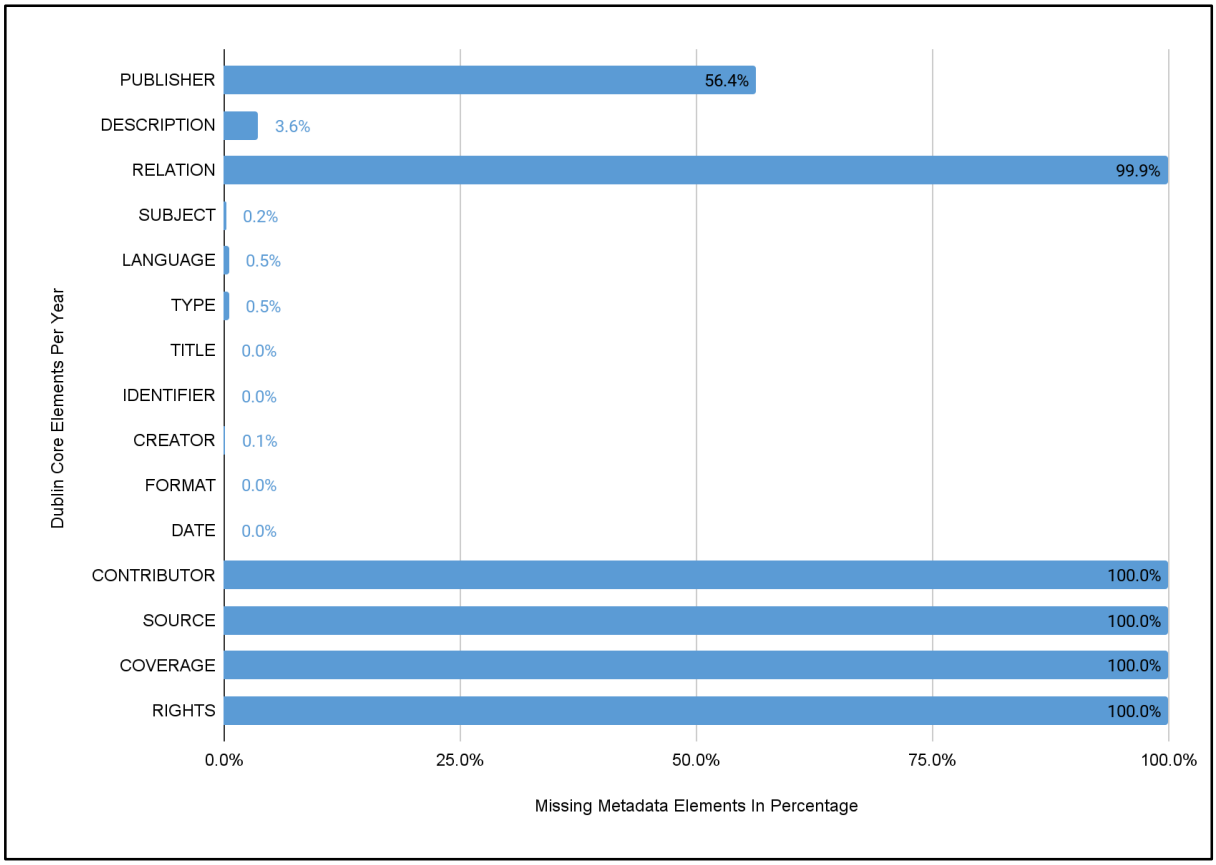


Figure 5.1.9: Number of missing yearly metadata elements

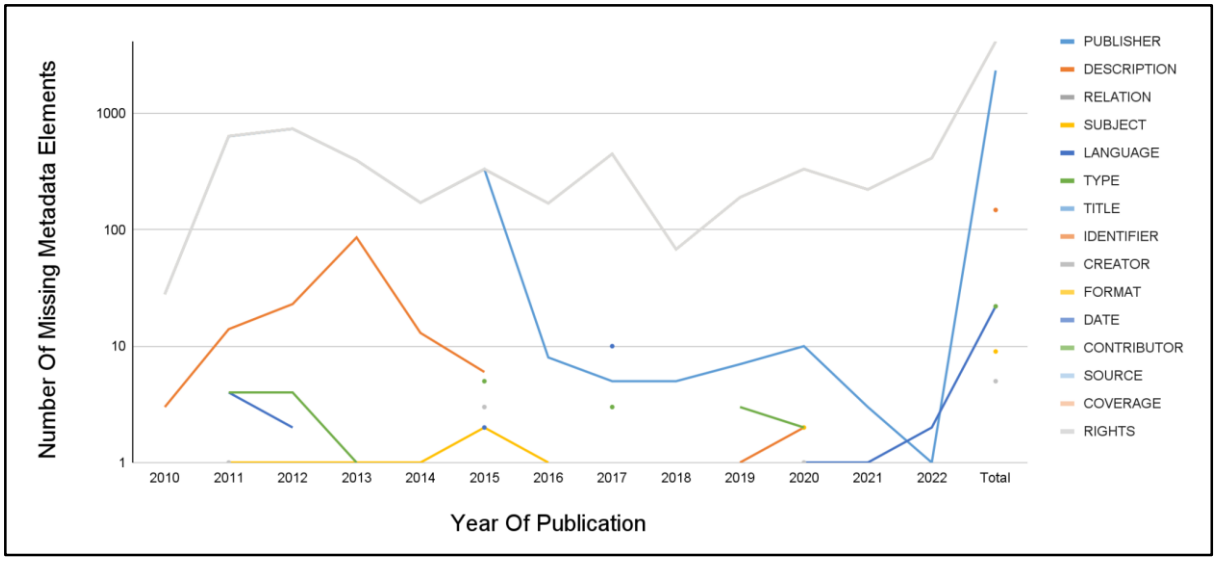


Figure 5.2.0: Graphical Line Representation of Yearly Missing Metadata Elements

5.1.3.2. Results for the cross match between OpenUCT vs UNZA repository

Comparatively, further data analysis was done between an exemplar record from OpenUCT as shown in figure 5.2.1 and the UNZA repository based on ETD-MS metadata elements shown in figure 2.2.0. From the analysis, it was observed that the exemplar record had a few missing metadata elements as compared to the UNZA repository that has a lot missing metadata elements as shown in Table 3.

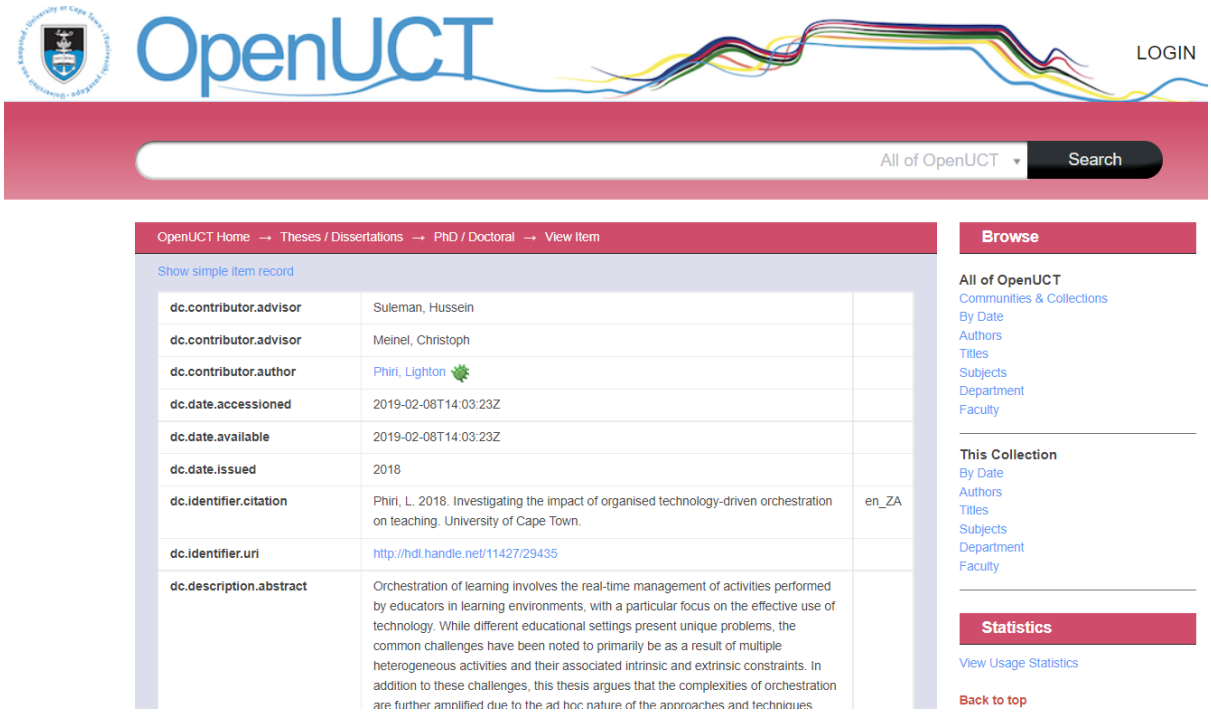


Figure 5.2.1: OpenUCT exemplar record

Table 3. Comparative Analysis; UNZA Vs. OpenUCT

ETD-MS Metadata Standard	**UNZA Repository**	**OpenUCT Repository**
Common Metadata Elements	**(4149) Records**	**(1) Exemplar Record**
dc.contributor	0	x
dc.contributor.advisor	0	x
dc.contributor.author	0	x
dc.contributor.role	0	x
dc.coverage	0	0
dc.creator	x	x
dc.date	x	x
dc.description.abstract	x	x
dc.description.note	0	0
dc.description.release	0	0
dc.publisher	x	0
dc.publisher.country	0	0
dc.publisher.institution	0	x
dc.publisher.faculty	0	x
dc.publisher.department	0	x
dc.relation	x	0
dc.rights	0	0
dc.subject	x	x
dc.title	x	x
dc.title.alternative	0	0
dc.type.qualification level	0	x
dc.type.qualification name	0	x
dc.type	x	x
dc.format	x	x
dc.identifier	x	x
dc.identifier.apa citation	0	x
dc.identifier.chicago citation	0	x
dc.identifier.vancouver citation	0	x

dc.identifier.ris	0	x
dc.language	x	x
thesis.degree	0	0
thesis.degree.name	0	0
thesis.degree.level	0	0
thesis.degree.discipline	0	0
thesis.degree.grantor	0	0
Total Number of missing metadata elements	**24**	**13**

5.2. Identification of the source of missing ETD metadata elements

The identification of the source of the Electronic Theses and Dissertations was done using an archival record analysis that involved 2 steps namely; the analysis of the DRGS guidelines and the sampling of records from all the different schools available on the UNZA repository.

5.2.1. Data Analysis

This involved a two-step approach.

Step 1. Analysing the DRGS postgraduate regulations to identify sections and how postgraduates are required to format the preliminary pages before submitting their ETDs to the university

In 2008, the Directorate of Research and Graduate Studies undertook a comprehensive review of the University Regulations for Postgraduate Studies which had been in use since 2006. A wide consultative process involving all internal stakeholders was adopted and completed in September 2009. The new regulations were approved by the Board of Graduate Studies in November the same year and officially came into effect on 20th July 2010. Since then, a number of areas have been identified which require strengthening thereby necessitating another review which also involved a consultative process resulting in the production of the March, 2015 version of the regulations.

These revised regulations were presented in the postgraduate regulations document in twenty nine self-explanatory sections and two appendices. The document derives strength from the University Act Number 11 of 1999, the University Senate and the General Regulations governing study at the University of Zambia.

Based on the new regulations for postgraduates, we identified the sections that guide postgraduates on how their Theses and Dissertations are supposed to be formatted, and the sections clearly showed the areas in the manuscripts where the missing metadata elements are located as illustrated in table 4.

Table 4. DRGS Regulation Analysis

DRGs SECTION	MANUSCRIPT SECTION	METADATA
Section 26	Title Page	Full Title and Subtitle
		Author's full name
		Publisher
		Date
	Copyright Declaration	rights
	Certificate of Approval	Author's full name
		Degree title
		Subject
		Supervisors (Contributors)
	Abstract	Abstract
	Language	Language
	Acknowledgements	Supervisor name (Contributor name)

Step 2. Sampling of 5 records from each school so as to identify locations of metadata elements on the pages

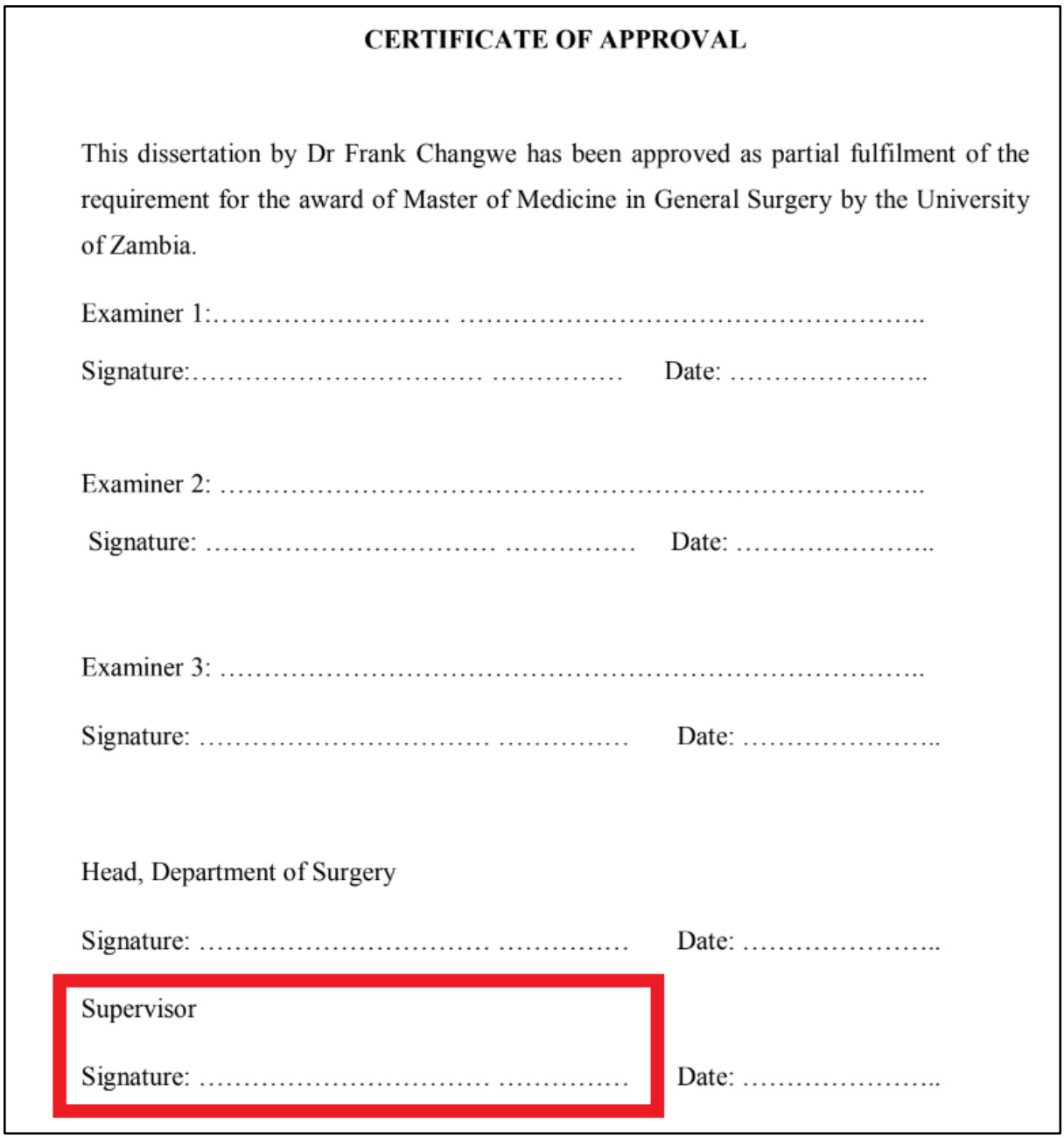
The University of Zambia has a total number of thirteen (13) schools, but it was observed that only a total number of twelve (12) schools have the uploaded theses and dissertations for postgraduate students. Therefore the random sampling of five (5) records was done from each school to identify the location of the metadata elements from the manuscripts. From the sampled records, the focus was on the preliminary sections of each document that contain the missing metadata elements. These sections include the title page, copyright declaration, certificate of approval, abstract, language and acknowledgement as shown in Table 7.A.1.

5.2.2. Results and further discussions

After undertaking the analysis of sixty (60) records from twelve (12) schools, the major outcomes of possible elements drawn from the analysis of twelve (12) schools showed that the metadata elements are found in the preliminary pages of the manuscripts. It was observed that the supervisor (contributor) details are found on the two sections namely, Approval section and Acknowledgement section of the manuscript. Unfortunately the certificate of approval in most documents only has a provision for the signature but does not have the contributor's details. Hence, the main focus of this research will be

based on the Acknowledgement section that contains contributor details from all the sampled documents.

Figure 5.2.2 shows the screenshot taken from one of the 60 sampling records from the approval section with the supervisor's signature and Figure 5.2.3 shows the acknowledgements with supervisors' details such as the names.



CERTIFICATE OF APPROVAL

This dissertation by Dr Frank Changwe has been approved as partial fulfilment of the requirement for the award of Master of Medicine in General Surgery by the University of Zambia.

Examiner 1:…… ……..

Signature:…… …… Date: ……..

Examiner 2: ……..

Signature: …… …… Date: ……..

Examiner 3: ……..

Signature: …… …… Date: ……..

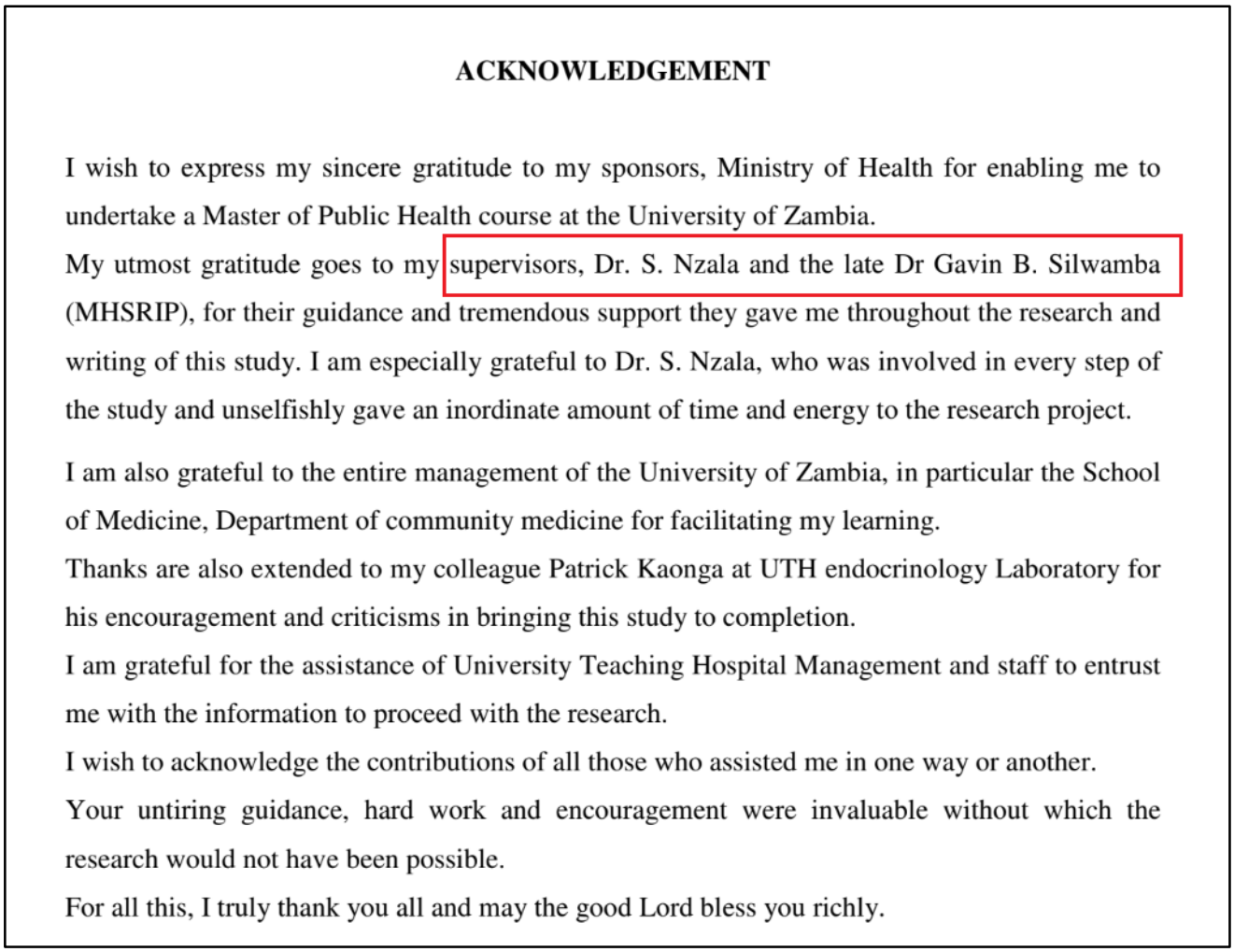
Head, Department of Surgery

Signature: …… …… Date: ……..

Supervisor

Signature: …… …… Date: ……..

Figure 5.2.2: Approval section with Supervisor's signature



ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my sponsors, Ministry of Health for enabling me to undertake a Master of Public Health course at the University of Zambia.

My utmost gratitude goes to my supervisors, Dr. S. Nzala and the late Dr Gavin B. Silwamba (MHSRIP), for their guidance and tremendous support they gave me throughout the research and writing of this study. I am especially grateful to Dr. S. Nzala, who was involved in every step of the study and unselfishly gave an inordinate amount of time and energy to the research project.

I am also grateful to the entire management of the University of Zambia, in particular the School of Medicine, Department of community medicine for facilitating my learning.

Thanks are also extended to my colleague Patrick Kaonga at UTH endocrinology Laboratory for his encouragement and criticisms in bringing this study to completion.

I am grateful for the assistance of University Teaching Hospital Management and staff to entrust me with the information to proceed with the research.

I wish to acknowledge the contributions of all those who assisted me in one way or another.

Your untiring guidance, hard work and encouragement were invaluable without which the research would not have been possible.

For all this, I truly thank you all and may the good Lord bless you richly.

Figure 5.2.3: Acknowledgements Section with the Supervisor's details

After analysing the 60 sample records from all the schools on the repository using the rule based technique as shown in Table 5, it was discovered that most of the records had the acknowledgement section on page number 6. Hence, our focus will be on page 6 from all the records uploaded on the repository from which the supervisor details will be extracted from using software techniques. Below is the table which shows the analysis of the records on the repository.

Table 5. Randomly Sampled Records' Acknowledgement Page Analysis

Page Number	Occurrences
7	20
6	27
4	2

9	1
5	5
8	4
2	1

5.3. Determination of appropriate extraction method of missing ETD elements from the PDF manuscripts.

The determination of the appropriate extraction method was done by using software techniques in three steps namely, downloading of the records from the repository, chopping off the acknowledgement sections from the manuscripts and lastly converting the section to a format that will enable easy extraction of the contributor details.

Step 1. Downloading of the dataset records

The process of downloading the 4149 records was done using Google colab by installing different environments like xmllint, then later on we came up with the shell script that had hard access to the Google drive directory where the records were kept after downloading. Lastly the bash script was used to download the dataset in a batch of 100 records as shown in Figure 5.2.4.

Figure 5.2.4: bash script

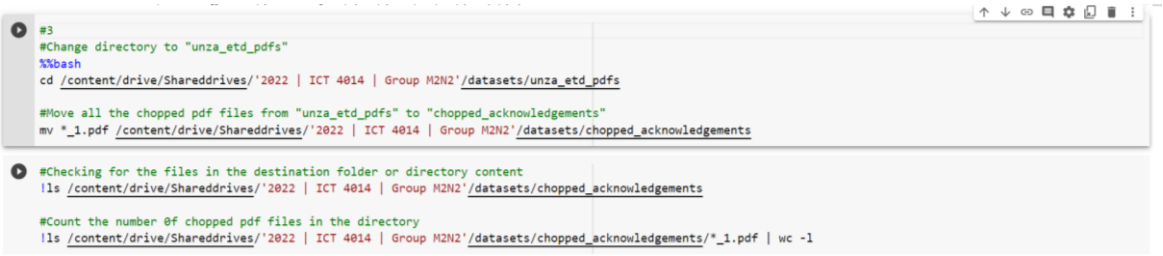
Step 2. Extraction of the Acknowledgement section from the manuscripts

After downloading all the records from the repository, page 6 of each document/record was chopped off using PyPDF2, converted to text and was put in one folder for easy access using a script. It was observed that, most of the ETDs were scanned documents and the script was set to skip the scanned records so as to avoid errors. Furthermore, the script was set to skip documents that had pages less than 6, this is because the basis was to only extract page six from all the manuscripts. Figure 5.2.5 shows the python script that was used to extract the acknowledgements, Figure 5.2.6 shows the script for

moving the extracted pages to the destination folder, Figure 5.2.7 shows the conversion script to text, and checking for the text files as shown in Figure 5.2.8. The extraction of the supervisor details was done by listing the text files in a directory using a python package namely glob, shown in Figure 5.2.7 and manipulating the data using python pandas as shown in Figure 5.3.0 and Figure 5.3.1.

```
#2
#Extract the Acknowledgement page(6) from multiple manuscripts
import PyPDF2
import os
import re
import sys
import glob
import PyPDF2 as pdf
from PyPDF2 import PdfFileReader, PdfFileWriter

path = glob.glob(os.path.join('/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_pdfs','*.pdf'))

for pdf_files in path:
  try:
    file_ext = pdf_files.replace('.pdf', '')
    pdf = PdfFileReader(pdf_files)
    pdfpage = [5]

    PdfWriter = PdfFileWriter() #Creating pdfWriter instance
    for page_num in pdfpage:
      PdfWriter.addPage(pdf.getPage(page_num))

    with open('{0}_1.pdf'.format(file_ext), 'wb') as a:
      PdfWriter.write(a)
      a.close()
  except Exception:
    pass
```

Figure 5.2.5: Python script for extracting the acknowledgement page



```
#3
#Change directory to "unza_etd_pdfs"
%%bash
cd /content/drive/Shareddrives/'2022 | ICT 4014 | Group M2N2'/datasets/unza_etd_pdfs

#Move all the chopped pdf files from "unza_etd_pdfs" to "chopped_acknowledgements"
mv *_1.pdf /content/drive/Shareddrives/'2022 | ICT 4014 | Group M2N2'/datasets/chopped_acknowledgements
```

```
#Checking for the files in the destination folder or directory content
!ls /content/drive/Shareddrives/'2022 | ICT 4014 | Group M2N2'/datasets/chopped_acknowledgements

#Count the number 0f chopped pdf files in the directory
!ls /content/drive/Shareddrives/'2022 | ICT 4014 | Group M2N2'/datasets/chopped_acknowledgements/*_1.pdf | wc -l
```

Figure 5.2.6: Python script for moving extracting pdf acknowledgement page

```
#5
######Convert all the extracted pdf page files in the directory "unza_etd_txts" to text########
import os
import PyPDF2

PDFS_FOLDER = '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/chopped_acknowledgements'
TEXTS_FOLDER = '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts'


def get_all_pdfs(folder_path):
      """
      :param folder_path: absolute folder path of the pdfs
      :return: a list with all the absolute path of pdfs
      """
      return os.listdir(folder_path)

def create_absolute_path(root_path, file_name):
      """
      :param root_path: absolute route path
      :param file_name: file name
      :return: absolute path of the file name
      """
      root_path = root_path + '/' if root_path[-1] != '/' else root_path
      return "%s%s" %(root_path, file_name)

def convert_pdf_to_text(pdf_path):
      """
```

Figure 5.2.7: Python script for converting extracted pdf acknowledgement pages to text

```
[ ] #Checking for the files in the destination folder or directory
    !ls /content/drive/Shareddrives/'2022 | ICT 4014 | Group M2N2'/datasets/unza_etd_txts/*_1.txt

    #Count the number @f converted pdf to text files in the directory
    !ls /content/drive/Shareddrives/'2022 | ICT 4014 | Group M2N2'/datasets/unza_etd_txts/*_1.txt | wc -l

    '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_1010_1.txt'
    '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_1034_1.txt'
    '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_1065_1.txt'
    '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_1190_1.txt'
    '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_1202_1.txt'
    '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_1268_1.txt'
```

Figure 5.2.8: python script for checking the generated text acknowledgement page

```
[ ] #installing pandas
    !pip install pandas
```

Multiple '.txt' Files Joining from GoogleDrive Folder; ** **Multiple sentences Joining to lines of strings for files;* * **Creating a DataFrame for all the files in pandas

```
[12] #importing pandas and glob
     import pandas as pd
     from glob import glob

     files = glob('/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts' + "/*.txt")
     files

     '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_7688_1 (1).txt',
     '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_7691_1 (1).txt',
     '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_7692_1 (1).txt',
     '/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts/123456789_7694_1 (1).txt',
```

Figure 5.2.9: importing pandas library and glob

```
[13] ##Joining multiple lines of text files into lines of strings(a list of strings)
     ##creating a spacy dictionary for the text files

     var_files_dict = {}
     for record in files:
       with open(record, 'r') as f:
         lines = f.readlines()
         jointsents = ' '.join(line.replace('\n', ' ') for line in lines)
         print(jointsents)
         print(type(record), record.split("/")[7])
         var_files_dict[record.split("/")[7]] = jointsents

     <class 'str'> 123456789_6621_1.txt
     v    DEDICATION    I dedicate this work to my dear husband and daughter, Monique Niza. You motivate me to work   even harder.
     <class 'str'> 123456789_6367_1.txt
     v    DEDICATION      I dedicate this w ork to my late father, Mr Joseph Mweni Chilungu Abunya Malembo , whose   inspiration and encouragements gave me the determination to reach thi
     <class 'str'> 123456789_6616_1.txt
```

Figure 5.3.0: python script for creating a dictionary for the text files

```
[14] ##creating a dataframe from the spacy dictionary
     finalDataFrame = pd.DataFrame([var_files_dict]).T
     finalDataFrame
```

1 to 25 of 1371 entries Filter

index	0
123456789_7176_1.txt	v ACKNOWLEDGMENT S I give praise and thanks to our God for giving me good health, understanding and many other blessings during the course of my study. I am extremely grateful to my Parents Mr and Mrs Samson Phiri for providing the support and encouragements throughout my study. I am extremely grateful to Dr. Phiri Jackson and Dr Lubobya Charles who helped enormously and guided me with their knowledge and experience throughout my research to completion. I would als o like to thank the Zambia Police Service for the support and also for providing me with all relevant information towards this project. Special thanks goes to UNZA Department of Computer Science and Department of Electrical and Electronic Engineering for t he advice and expertise rendered to this project. Last but not the least, I would like to thank all my course mates not only for the wonderful time and networks we have created but also for the support and encouragements rendered to me, these are; Benaiah Akombwa, Mathias Kamanga, Thomas Muyumba, Thandiwe Mphande and Winter Musukwa.
123456789_7053_1.txt	DEDICATION I dedicate this work to my parents, Annette Mutale Mulenga and Moses Mulenga, who are an epitome of excellent parenthood and without whom I would not have achieved anything academically. Their unwavering spiritual, emotional and financial support is much appreciated. v
123456789_7034_1.txt	vi ACKNOWLEDGEMENTS The idea and motivation for this work emerged majorly from a lecture delivered at the University of Zambia's School of Engineering entitled "Economics of Distributed Resources". Zambia , the country of analysis is of special interest to me as my domain country that belongs to the emerging countries whose econ omic development is driven by wood fuel, hydropower save for the depletion of fossil resources. Given today's carbon -constrained world, it was impera tive to evaluate the competitiveness of alternative energy sources. The research wouldn't have been made possible without the valuable professional, moral and financial support from others. First and foremost, I take the opportunity to thank my supervisor/lecturer Dr. A Zulu for the valuable academic guidance and inspiring brain - storming aside discussions we h ad together, which eventually helped me research into this interesting topic. Furthermore, I would like to give special thanks to all my lecturers at UNZA especially Prof. F Yamba, Dr. H Mwenda, Dr. L Ngoyi, Dr Simate, Dr. M Chileshe, and Mr Virdy, for th eir excellent academic support and guidance. Moreover, I am very grateful to Prof. Wilfried Zorner (AIR Project Leader, Head of InEs, Professor/Senior Lecturer at THI), Prof. D Navarro (THI Professor/Senior Lecturer), Ms. Linda Ehrl (AIR Project Coordinat or at InEs/THI), Dr. Christoph Trinkl, Mr Stefan, Mr Sebastian Sommer, Mr Abdessamad Saidi and Mr Matthias (Buddies at THI/InEs) for their great support during my AIR Student Exchange visit to THI, InEs and Technology Excursions in Germany. During my study , I had the privilege to benefit from the AIR Student Exchange Scholarship for Renewable Energy, which enabled me focus on this research thesis . Finally, I am very grateful for all the support, discussions and feedback from my family, friends and the interface with my fellow students who helped me along the way. Above all, I would like to extend my special thanks to my wife Diana Longwe Mwaala for her enduring support and encouragement throughout the work.
123456789_7001_1.txt	v ACKNOWLEDGEMENTS I would like to express my sincere gratitude to my supervisor Mr. Thomas Mtonga. This work would not have been what it is without your continual and tireless supervision. Your comments were invaluable at all times. Besides, I would like to thank Dr. Elbers Willem and Liliane Foundation for their support. Further gratitude goes to Dr. Daniel Ndhlovu, Dr. Joseph M. Mandyata, Dr. Beatrice Matafwali, Dr. Peter Chom ba Manchishi and Dr. M agdalene . Simalalo for their powerful lectures. I thank my course mates for the support and memorable moments shared. Kabaka Musonda in particular, your encouragement made me work harder. Last but not the least, I would like to thank my husband Fanizani Phiri and my children Idah, Penias, Thungela and Alyssa for their support, encouragement and understanding during the entire course of my study. Glory to God Almighty for His grace, love and mercies during the course of this study.
123456789_6996_1.txt	v ACKNOWLEDGMENTS I would like to extend my gratitude to my brother Clifford Hepplethwaite, without whom I would not be where I am today, I am grateful for all the support and encouragement that he gave me throughout this study. I would like to thank my other family members and friends for their continuous support. I would like to extend my thank s to my supervisor Dr. Bridget Bwalya Umar, for constantly believing in me and pushing me to do my best, for the guidance and wisdom that she continuously showed me. I would also like to say thank you to all the lecturers and staff at the University of Zambia, School of Natural Sciences, Geography Department, who have taught me a lot and made me a better person overall.

Figure 5.3.1: Python pandas DataFrame for the text files

Step 3. Extraction of Contributor's details from the Acknowledgement section

Extraction of the supervisor's details was done using spacy built-in functionalities that are capable of identifying the figure of speech from the document's sentences. These functionalities include name recognition functionality using prescribed entities (PERSON, ORG) as shown in Figure 5.3.2 below. By the use of basic string matching, the sentences containing the specific phrase matching words were extracted and the supervisor's details were identified from the sentences. It was observed that while trying to extract the supervisor details from the acknowledgements, the software library leaves out the salutation for the names. This is because SpaCy has a pre-set figure of speech that is capable of identifying the name from the sentences. In addition to that, records that never had the supervisor details were automatically skipped by the script. Figure 5.3.3, Figure 5.3.4, Figure 5.3.5 and Figure 5.3.6 show the various procedures that were taken to extract the supervisor details from the text files.

Figure 5.3.2: example individual record's spacy named entities and their labels

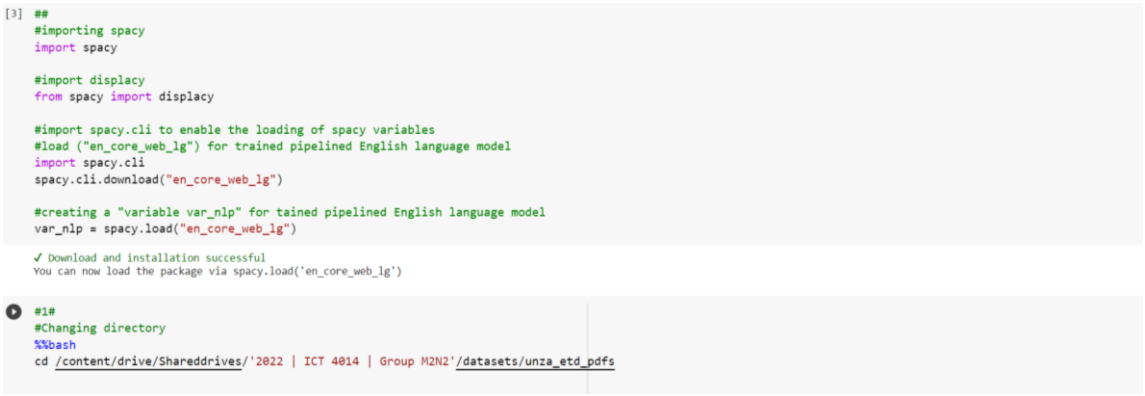


Figure 5.3.3: importing SpaCy library and its dependencies

```
#2#
#####Basic string matching
#####printing sentences with UpperCase/LowerCase "supervisor" or "co-supervisor"
#####create function for the output

import spacy
from spacy.matcher import PhraseMatcher

def var_function(var_filename):

  with open(var_filename, 'r') as f:
      lines = f.readlines()
      text = ' '.join(line.replace('\n', ' ') for line in lines)

  var_ict4014_example = text

  var_ict4014_example_doc = var_nlp(var_ict4014_example)

  #call on the spacy document containing the joined lines
  text = var_ict4014_example_doc

#phrase matching and pattern identification
  phrase_matcher = PhraseMatcher(var_nlp.vocab)
  phrases = ['Supervisor', 'co-supervisor', 'supervise','supervisors', 'Co-supervisor', 'supervisor']
  patterns = [var_nlp(text) for text in phrases]
  phrase_matcher.add('', None, *patterns)

  doc = var_nlp(text)
  var_supervisors_list = []
#check for the predefined matching phrases from the sentences
  for sent in doc.sents:
      for match_id, start, end in phrase_matcher(var_nlp(sent.text)):
          if var_nlp.vocab.strings[match_id] in [""]:
#             print(sent.text, type(sent.text))
```

Figure 5.3.4: python script for basic string matching and supervisor identification

```
#3#
#print the identifier alligned with supervisor details seperated by the a full colon

import os
import pandas as pd
from glob import glob

files = glob('/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts' + "/*.txt")

for var_file in files:
  print(os.path.basename(var_file), ":", var_function(var_file))
```

```
123456789_6222_1.txt :
M. Makasa
Busiku Hamainza
123456789_6223_1.txt : M. Makasa=Busiku Hamainza
Samuel Bosomprah
Chris Mwe emba
Wilbroad Mutale
```

Figure 5.3.5: python script for extracting supervisor aligned with the record identifier

```
[6] #3#
    #Extract the multiple records' metadata and save the output as an excel file
    #!/usr/bin/env bash

    import os
    import pandas as pd
    from glob import glob

    files = glob('/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/datasets/unza_etd_txts' + "/*.txt")

    for var_file in files:
      print(os.path.basename(var_file), ":", var_function(var_file), file=open("unza_etds.xlsx", "a"))
```

```
Ravi P aul
Dr  Tommie N
Dr Hambulo
Rose   Chikopela
Joseph Mandyata
Mwiza   Nakamba
Moses Chanzala
```

Figure 5.3.6: python script for writing the extracted supervisors details to an Excel file

Step 4. Loading the extracted contributor's details into a pandas dataframe

After extracting specific phrases matching words for the supervisor's details, they were then saved in a named entity for easy access when loading them into a pandas dataframe created using a library spacy dictionary imported as pandas pd. Details loaded on the data frame had to be the same as those in the columns of the spreadsheet for the ground truth. Figure 5.3.7 shows the machine generated supervisors

from the respective records identified by the unique identifier from the UNZA repository. The University of Zambia carried out a manual transcription of records that were available in the repository, this transcription enabled the collection of the metadata elements for each record in the repository. The transcription was based on the set rules on how the tasked individuals were required to collect the metadata elements from the manuscripts. In this research, we used the transcription data as the ground truth for comparing the machine generated to and came up with the comparative analysis between the two data sets as shown in Figure 5.2.9.

	automatically generated supervisor
2	123456789_7176_1.txt :
3	123456789_7053_1.txt :
4	123456789_7034_1.txt :
5	123456789_7001_1.txt : Thomas Mtonga
6	123456789_6996_1.txt : Bridget Bwalya Umar
7	123456789_6998_1.txt :
8	123456789_7029_1.txt : C.M. Namafe
9	123456789_7028_1.txt :
10	123456789_7022_1.txt :
11	123456789_7020_1.txt :
12	123456789_6997_1.txt : Oswell C. Chakulimba=H. Mbozi=Oswell C. Chakulimba=H. Mbozi
13	123456789_7003_1.txt : Dr Dennis Banda
14	123456789_7002_1.txt :
15	123456789_6995_1.txt :
16	123456789_7000_1.txt : Wanga W
17	123456789_6985_1.txt :
18	123456789_7122_1.txt : Balimu Mwiya
19	123456789_7033_1.txt :
20	123456789_7032_1.txt :

Figure 5.3.7: Automatically generated supervisor

	identifier	dc.contributor.advisor
2	123456789_3777	S. KASONDE NG'ANDU= .E. MUNSAKA
3	123456789_4729	DANIEL NDHLOVU
4	123456789_3233	SOPHIE KASONDE-NG'ANDU
5	123456789_980	
6	123456789_2198	
7	123456789_1997	P.C MACHISHI
8	123456789_798	C MICHELO
9	123456789_4728	SUMBUKENI KOWA
10	123456789_4750	LT MUUNGO = N LAMBWE
11	123456789_1310	M.P.S NGOMA= B.C.AMADI
12	123456789_4086	G. TEMBO= .T. KALINDA
13	123456789_334	HENRY J. MSANGE
14	123456789_4870	JAMES MUNTHALI
15	123456789_3731	CHIPEPO KANKASA = ELWYN CHOMBA
16	123456789_3745	MAFULEKA,
17	123456789_4302	KASONDE-NG'ANDU
18	123456789_1613	K.S. BABOO
19	123456789_4201	W.S. KALIKITI
20	123456789_3821	P.TEMBO, E.T.ODIMBA= MUNTHALI

Figure 5.3.8: Ground Truth

```
[7] #import the ground truth from the google drive directory as a pandas dataframe
    import pandas as pd
    data = pd.read_excel(r'/content/drive/Shareddrives/2022 | ICT 4014 | Group M2N2/research_report/Project Evaluation/final_evaluation.xlsx')
    #display(pd.DataFrame(data))
    df = pd.DataFrame(data)
    df
```

1 to 25 of 77 entries Filter

index	identifier	automatically.generated.supervisor	ground_truth
0	123456789_4808	SUMBWANYAMBE	SUMBWANYAMBE
1	123456789_4876	M. C. M. BWALYA	M. C. M. BWALYA
2	123456789_5562	ZALI IAN	BANDA NZILA
3	123456789_4844	MUTALE W. CHANDA=RADHE KRISHNA	MUTALE W. CHANDA= RADHE KRISHNA
4	123456789_3767	C. M.	C.M. NAMAFE
5	123456789_4915	R.N. LIKWA=J. BANDA	R. N. LIKWA= J. BANDA
6	123456789_4263	JOSEPH NG'NDU	R JOSEPH NG'NDU
7	123456789_3854	H	HENRY. J. MSANGO
8	123456789_2716	KENNY MAKUNGU	NaN
9	123456789_3651	I. A. NYAMBE	I. A. NYAMBE
10	123456789_3408	VICTOR SHIKAPUTO=NUM	VICTOR SHIKAPUTO
11	123456789_4106	FIDELIS MUZYAMBA	FIDELIS MUZYAMBA
12	123456789_2971	S. WA SOMWE=B. AMADI	S. WA SOMWE= B. AMADI
13	123456789_3031	M.Y=M. LEMBA	M. LEMBA
14	123456789_3032	W. CHAKANIKA=A NOLT=L. H. MOONGA=E MMY MBOZI=NOAH K. SICHULA	WANGA W. CHAKANIKA
15	123456789_3052	A. AKAKANDELWA	A. AKAKANDELWA
16	123456789_3430	D.M. LUNGU	D.M. LUNGU
17	123456789_3011	BANDA DENNIS	DENNIS BANDA
18	123456789_4135	MB	CHISHIMBA
19	123456789_3650	WANGA W. CHAKANIKA=PATRICK S. NGOMA=ANOLT L. H. MOONGA=EMMY MBOZI=NOAH K. SICHULA=ROICK CHONGO	WANGA W. CHAKANIKA
20	123456789_4109	R. TEMBO=B. NSEMUKOLA=TIM E.	R. TEMBO= B. NSEMUKOLA

Figure 5.3.9: loaded DataFrame for the automatically generated supervisor Vs. ground truth

CHAPTER 6

6. System Evaluation

6.1. Automated generation vs Ground truth

The evaluation of the effectiveness of the automatic generation of the missing metadata involved creating two datasets namely, the excel spreadsheet for the automatically generated supervisor details in figure 27 and Google spreadsheet for the ground truth in figure 28 extracted from the transcription that was done. Furthermore, a pandas dataframe shown in figure 29 was created from the two spreadsheets so as to have a clear comparative analysis using Natural Language evaluation metrics.

6.2. Selected Natural Language Evaluation Metrics

There is a good number of Natural Language evaluation metrics for any given model that can be used to evaluate the likeness of the machine generated texts. Below are the metrics we tried and tested for machine generated results

6.2.1. BLEU (Bilingual Evaluation Understudy)

This is a metric for automatically evaluating machine-translated text. It is a performance metric to measure the performance of machine translation models. It evaluates how well a model translates from one language to another. It assigns a score for machine translation based on the unigrams, bigrams or trigrams present in the generated output and compares it with the ground truth [4]. It has many problems but it was one of the first methods to assign a score to machine translation models. It always gives a score between 0 and 1. The BLEU score measures the similarity of the machine-translated text to a set of high quality reference translations, where one represents 100% similarity and zero represents 0% similarity. Figure 6.1.0 shows the predefined formula for BLUE and Figure 6.2.0 shows the example of BLUE's identified n-gram.

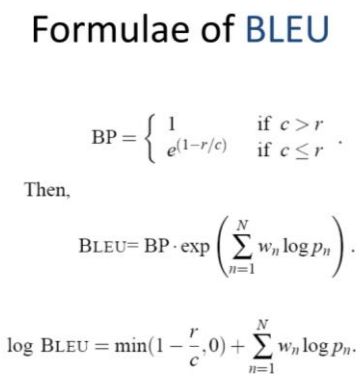


Figure 6.1.0: BLEU predefined formula

unigram	bigram	trigram
Once	Once you	Once you stop
you	you stop	you stop learning
stop	stop learning	stop learning, you
learning	learning you	learning, you start
you	you start	you start dying
start	start dying	
dying		

Figure 6.2.0: Examples of BLEU's identified n-gram

6.2.2. *ROUGE (Recall-Oriented Understudy for Gisting Evaluation)*

This is Recall based, unlike BLEU which is Precision based. This is a set of metrics used for evaluating automatic summarization and machine translation software in NLP [11]. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

5.3. Testing Natural Language Evaluation Metrics

In order to try out the performance of each of the metrics being used in the research project model, the N-gram method test was used in order to measure the performance of the metrics. The procedure was done in two ways: the first part is randomly choosing an index number for a supervisor's name on the

dataset loaded on the pandas dataframe. The second part was to specify N-gram occurrences associated with the supervisor name for validating and testing the model.

To evaluate the performance of the model, numerical matrix, which is used to measure the quality of the generated supervisors' names by comparing its closeness to one or more ground truth human inputs e.g. a translation using the same words (1-grams) as in the references tends to satisfy adequacy. The longer n-gram matches account for fluency. The performance measure metrics used were the Precision and Recall. These two metrics are defined below as follows:

Precision

This counts up the number of candidate translation (human imputed supervisor names) words (unigrams) which occur in any reference translation (ground truth) and then divides by the total number of words in the candidate translation (generated supervisor names). Precision, therefore, calculates the accuracy for the minority class. The formula for precision is as follows:

$$Countclip = min\ (Count, Max\ Ref\ Count).$$

n-gram precision on a multi-sentence test set:

$$pn = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Countclip\ (n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Countclip\ (n\text{-}gram')}$$

Recall

This measures the candidate summary and a set of reference summaries. This is computed as follows:

$$Countmatch = min\ (Max\ candidate\ Count, Ref\ summaries).$$

$$Rn = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{n\text{-}gram \in S} Countclip(n\text{-}gram)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{n\text{-}gram \in S} Countclip(n\text{-}gram)}$$

The trouble with recall

Traditionally, precision has been paired with recall to overcome such length-related problems. However, BLEU considers multiple reference translations, each of which may use a different word choice to translate the same source word. Furthermore, a good candidate translation will only use (recall) one of these possible choices, but not all. Indeed, recalling all choices leads to a bad translation. Here is an example:

Candidate 1: I always invariably perpetually do. **Candidate 2**: I always do. **Reference 1**: I always do. **Reference 2**: I invariably do. **Reference 3**: I perpetually do.

The first candidate recalls more words from the references, but is obviously a poorer translation than the second candidate. Thus, naive recall computed over the set of all reference words is not a good measure. Admittedly, one could align the reference translations to discover synonymous words and compute recall on concepts rather than words. But, given that reference translations vary in length and differ in word order and syntax, such a computation is complicated.

5.4. The BLEU Evaluation

The BLEU metric ranges from 0 to 1. In this research project, we worked with 77 BLEU scores that were generated from the records that were extracted from the pdf manuscripts that were used as the datasets shown in Table 8.A.2. Based on the metrics, few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. It is important to note that the more reference translations per sentence there are, the higher the score is. Thus, one must be cautious making even "rough" comparisons on evaluations with different numbers of reference translations. BLEU is precision based in the sense that the numerator contains the sum of the overlapping □-grams across all the hypotheses (i.e., all the test instances) and the denominator contains the sum of the total □-grams across all the hypotheses (i.e., all the test instances). Here, each □□□□□□□□□□ is summed over all the hypotheses, thus, BLEU is called a corpus-level metric, i.e. BLEU gives a score over the entire corpus (as opposed to scoring individual sentences and then taking an average).

Based on the above methods and procedures undertaken in testing the metrics, the overall evaluation for the Rule-based automatic generation of missing ETDs metadata was settled on using the BLEU (*Bilingual Evaluation Understudy*) because it rapidly corresponded to effective modelling ideas. BLEU is a quality metric score for machine translation systems that attempts to measure the correspondence between a machine translation output and a human translation. Figure 6.4.0 below shows the results for

automatic generated results obtained after using BLEU compared to the ground truth. Figure 6.5.0 and Figure 6.6 shows the graphical representation of the BLUE scores.

Measuring translation quality is difficult because there is no absolute way to measure how “correct” a translation is. Machine translation is a particularly difficult AI challenge because computers prefer binary outcomes, and translation has rarely if ever only one single correct outcome [15]. Many “correct” answers are possible, and there can be as many “correct” answers as there are translators. The most common way to measure quality is to compare the output strings of automated translation to a human translation text string of the same sentence. The fact that one human translator will translate a sentence in a significantly different way than another human translator, leads to problems when using these human references to measure “the quality” of an automated translation solution. Particularly, the evaluation of this research’s BLEU scores makes reference to the transcription that was human translated by individuals that were assigned to manually identify the supervisor details from the pdf manuscripts, the generated scores show good machine translation when it comes to the accuracy as they generate out at corpus level.
The interpretation of the BLEU scores is based on the frequency ranges (0.25 intervals) of the scores. From our results, we can clearly see that there were multiple variations, this is because of the mismatch between ground truth value and the solution value and this was caused by the following reasons. Firstly, the ground truth was human generated and so human errors are possible while generating the values. Secondly, this research made use of BLEU’s bigrams to generate possible scores for the most usual two names that are used as first name and last names for people in the official documentations and that simply meant that each value for first name in the solution was compared to exactly one value for first name in the ground truth, which also applies to the last names. Hence if the first name was swapped or left out or misspelled in the ground truth, then automatically the match led to a zero or lower score. Observably, from the scores in the range between 0.0 and 0.25 scores there is quite a small number of them, the scores in the range between 0.25 and 0.5 scores are little, which also applies to the scores between 0.5 and 0.75 scores. Conversely, we clearly see that scores between 0.75 and 1.00 had a bigger number and this is because of the precise matching of the ground truth values and the solution values.

```
[['SUMBWANYAMBE']] ['SUMBWANYAMBE']
1.0
[['M.', 'C.', 'M.', 'BWALYA']] ['M.', 'C.', 'M.', 'BWALYA']
1.0
[['BANDA', 'NZILA']] ['ZALI', 'IAN']
0
[['MUTALE', 'W.', 'CHANDA', 'RADHE', 'KRISHNA']] ['MUTALE', 'W.', 'CHANDA', 'RADHE', 'KRISHNA']
1.0
[['C.M.', 'NAMAFE']] ['C', '.M.']
0
[['R.', 'N.', 'LIKWA', 'J.', 'BANDA']] ['R.N.', 'LIKWA', 'J.', 'BANDA']
0.5841005873035536
[['R', 'JOSEPH', "NG'NDU"]] ['JOSEPH', 'NG"NDU']
0.3032653298563167
[['HENRY.', 'J.', 'MSANGO']] ['H']
0
[[]] ['KENNY', 'MAKUNGU']
0
[['I.', 'A.', 'NYAMBE']] ['I.', 'A.', 'NYAMBE']
1.0
[['VICTOR', 'SHIKAPUTO']] ['VICTOR', 'SHIKAPUTO', 'NUM']
0.6666666666666666
[['FIDELIS', 'MUZYAMBA']] ['FIDELIS', 'MUZYAMBA']
1.0
[['S.', 'WA', 'SOMWE', 'B.', 'AMADI']] ['S.', 'WA', 'SOMWE', 'B.', 'AMADI']
1.0
[['M.', 'LEMBA']] ['M', 'Y', 'M.', 'LEMBA']
0.5
[['WANGA', 'W.', 'CHAKANIKA']] ['W.', 'CHAKANIKA', 'A', 'NOLT', 'L.', 'H.', 'MOONGA', 'E', 'MMY', 'MBOZI', 'NOAH', 'K.', 'SICHULA']
0.15384615384615385
[['A.', 'AKAKANDELWA']] ['A.', 'AKAKANDELWA']
1.0
[['D.M.', 'LUNGU']] ['D.M.', 'LUNGU']
```

Figure 6.4.0: BLEU scores for automatically generated results Vs. ground truth

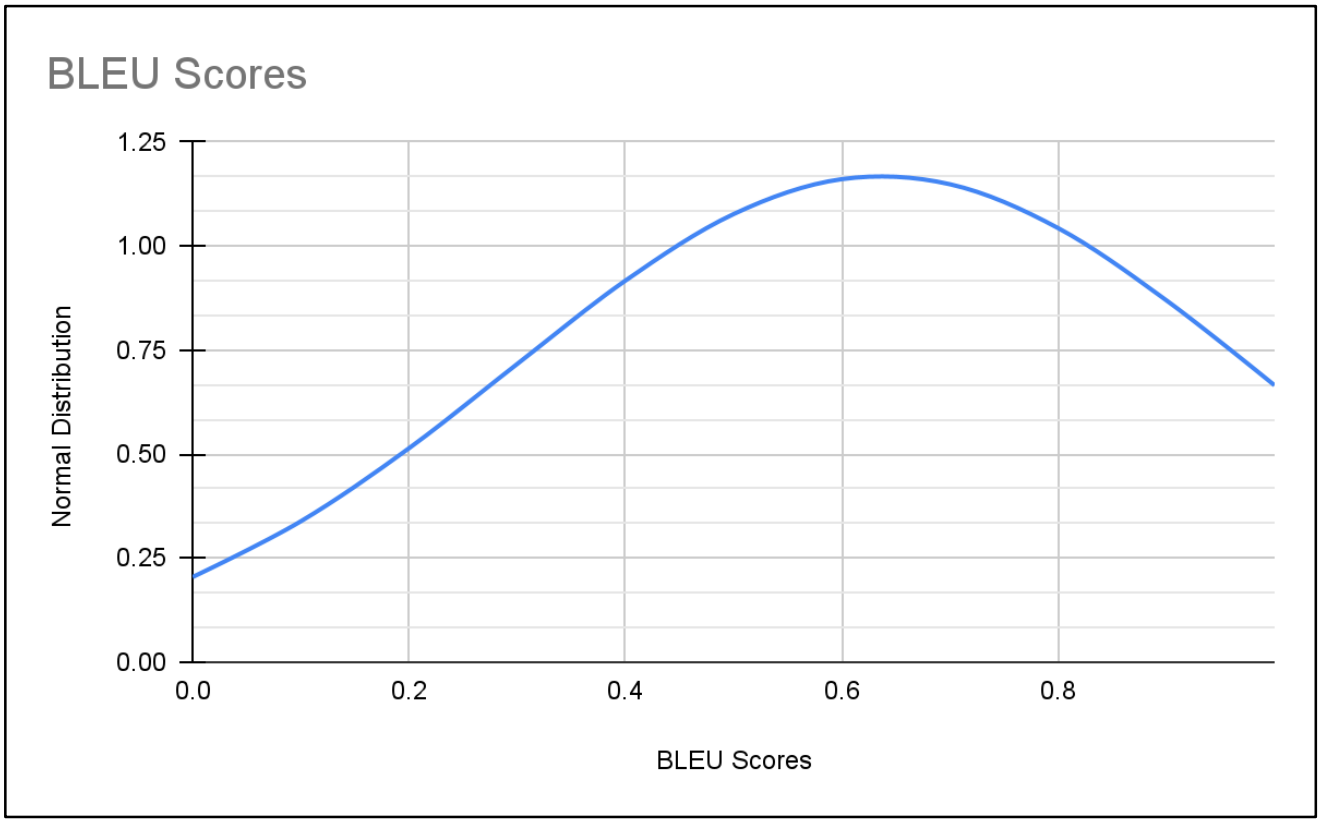


Figure 6.5.0: graphical representation of BLEU scores using a bell curve

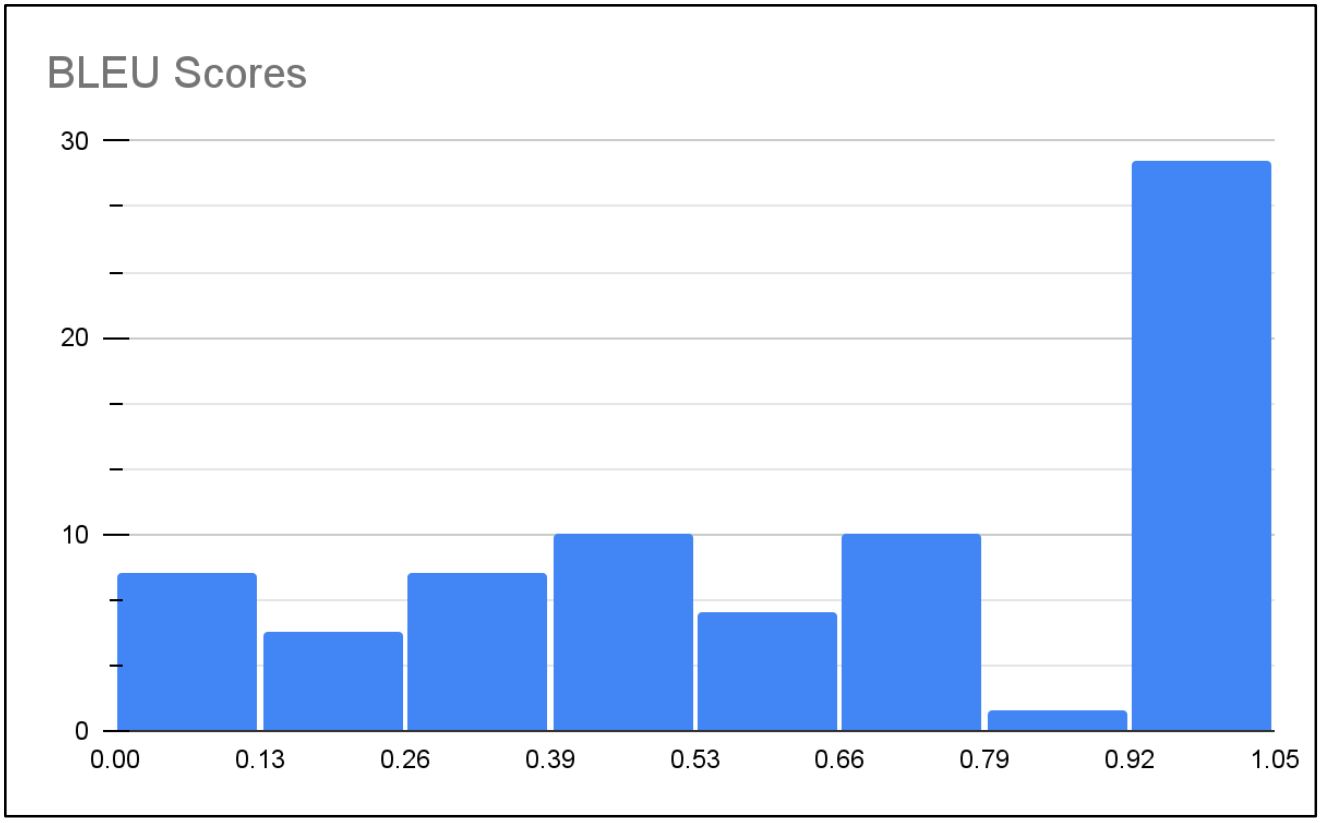


Figure 6.6.0: graphical representation of BLEU scores using a histogram

7. Conclusion

Automatic generation of electronic theses and dissertation metadata plays an important role in the searching of the particular record metadata by multiple researchers. This solution has been tried and tested using natural language processing metrics models machine generated against human generated results at individual record level, and it has been proven to be the suitable solution for generating metadata elements provided that the prescribed document formatting is followed by researchers when writing their research output. Particularly, BLEU as a precision based metric was chosen as an evaluation metric to evaluate the effectiveness of the automatically generated supervisors when compared to the human generated results. This solution mainly makes use of software libraries that are free and open source, which simply means that users wouldn't incur any charges for using them in the process of automatically generating metadata elements. This research's relevance is depicted in the automatic retrieval of research output before uploaded to some downstream services and in the referencing. Based on the research that was conducted, the main findings include the following. Identification of the missing metadata elements from the UNZA repository only has 10 available metadata that also lacked consistency in the association of the metadata elements and 5 other metadata elements were totally missing from the UNZA repository. After the previous research study findings, the identification of the source of the missing metadata gave results that showed that the supervisor details are located on the acknowledgements section of the manuscripts. The last study on the determination of the appropriate extraction method resulted in the use of a rule based approach which made use of software libraries like Spacy. The evaluation of this machine generated results gave a good

overview of the output with reference to the manually human generated metadata. In addition to that, this extraction of metadata elements was tested on individual record and on corpus level using the bigrams of BLEU. This research will change the way people have access to metadata elements from different records, so instead of reading through the whole document to get the supervisor's name or title name, these entities can be automatically generated from the documents which makes life easy and less time consuming. This will also help users out there to know the documents which are of high quality metadata in line with which documents were scanned or typed before being uploaded to the repository.

References

[1] Chipangila, B., Liswaniso, E., Mawila, A., Mwanza, P., Nawila, D., M'sendo, R., Nyirenda, M. and Phiri, L. 2021. Improved Discoverability of Digital Objects in Institutional Repositories Using Controlled Vocabularies. *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

[2] Dublin Core™ Metadata Element Set, Version 1.1: Reference Description: *https://www.dublincore.org/specifications/dublin-core/dces/*. Accessed: 2022-11-17.

[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#*. Accessed: 2022-11-17.

[4] Evaluating models: *https://cloud.google.com/translate/automl/docs/evaluate*. Accessed: 2022-11-16.

[5] Gunjan, V.K. and Zurada, J.M. 2021. *Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough: Latest Trends in AI, Volume 2*. Springer Nature.

[6] Introduction to Named Entity Recognition: *https://www.kdnuggets.com/introduction-to-named-entity-recognition.html*. Accessed: 2022-11-18.

[7] 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*.

[8] Phiri, L. 2018. Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia. *IEEE International Conference in Information and Communication Technologies*.

[9] PyPDF2 Library for Working with PDF Files in Python: 2021. *https://www.analyticsvidhya.com/blog/2021/09/pypdf2-library-for-working-with-pdf-files-in-python/*. Accessed: 2022-11-18.

[10] Qiu, S. and Zhou, T. 2019. A method of extracting metadata information in digital books. *2019 10th International Conference on Information Technology in Medicine and Education (ITME)* (Aug. 2019).

[11] rouge-metric: *https://pypi.org/project/rouge-metric/*. Accessed: 2022-11-16.

[12] Smalheiser, N.R. and Torvik, V.I. 2009. Author name disambiguation. *Annual Review of Information Science and Technology*.

[13] Steward, S. 2004. *PDF Hacks: 100 Industrial-Strength Tips & Tools*. "O'Reilly Media, Inc."

[14] Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P.J. and Bolikowski, Ł. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*.

[15] Towards Hybrid Human-Machine Translation Services: 2018. *https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjrhvu18Lb7AhVPi1wKHTsGBY4QFnoECAgQAQ&url=https%3A%2F%2Fwww.dfki.de%2Ffileadmin%2Fuser_upload%2Fimport%2F9879_CI_2018_paper_22_%25283%2529.pdf&usg=AOvVaw1-uMMO8yQaqFKOp60BMImu*. Accessed: 2022-11-18.

[16] Vasiliev, Y. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press.

[17] Witt, M. 2010. An introduction to ORE. *Library technology reports*. 46, 4 (2010), 5–11.

[18] 1996. *Research Design: Qualitative and Quantitative Approaches*. John W. Creswell. *The Library Quarterly*.

8. Appendix A: Raw Results

Table 6. A.1: Acknowledgements Page Analysis for the 60 Random Samples

NAME OF SCHOOL	THESES/DISSERTATION AUTHOR	URL	ACKNOWLEDGEMENT PAGE NUMBER
Engineering	MUTINTA MATAMBO	http://dspace.unza.zm/handle/123456789/7519	7
	CHAMA CHILOMO	http://dspace.unza.zm/handle/123456789/7271	7
	MARY MABO NYAYWA	http://dspace.unza.zm/handle/123456789/7195	6
	JORDAN ZIMBA	http://dspace.unza.zm/handle/123456789/7019	6
	JONATHAN PHIRI	http://dspace.unza.zm/handle/123456789/7176	6
Education	FRIDAH MULENGA CHILUFYA	http://dspace.unza.zm/handle/123456789/7685	6
	INDIE KAYAMA	http://dspace.unza.zm/handle/123456789/7684	6
	COLLINS MILUPI	http://dspace.unza.zm/handle/123456789/7682	7
	BY BIGGIE CHANDA	http://dspace.unza.zm/handle/123456789/7681	6
	BENNY KAPOMPO	http://dspace.unza.zm/handle/123456789/7683	4
Library	LOMBE CHILESHE	http://dspace.unza.zm/handle/123456789/7515	6
	BESTAIN HAMPWAYE	http://dspace.unza.zm/handle/123456789/7350	7
	INONGE IMASIKU	http://dspace.unza.zm/handle/123456789/6228	7
	CHILESHE GREGORY	http://dspace.unza.zm/handle/123456789/6141	7
	SARAH GWAYI	http://dspace.unza.zm/handle/123456789/5107	7

Law	STEVEN NYUNDO	http://dspace.unza.zm/handle/123456789/6713	6
	SHUBAYI CHATORA	http://dspace.unza.zm/handle/123456789/6380	7
	SARA MULWANDA	http://dspace.unza.zm/handle/123456789/2346	9
	NANCY CHEWE MULENGA	http://dspace.unza.zm/handle/123456789/6247	7
	JENNIPHER BWALYA	http://dspace.unza.zm/handle/123456789/6185	7
Institute Of Distance Education	SIMACHELA KUWA	http://dspace.unza.zm/handle/123456789/7679	6
	MABLE N NAMOOBE	http://dspace.unza.zm/handle/123456789/4947	7
	CHIYEZHI JONATHAN	http://dspace.unza.zm/handle/123456789/5892	6
	CHILIMA, PHILLIP	http://dspace.unza.zm/handle/123456789/5901	5
	CHITALU JULIUS MUSONDA	http://dspace.unza.zm/handle/123456789/7395	4
Humanities and Social Sciences	FELISTUS MANGUNGA	http://dspace.unza.zm/handle/123456789/7282	8
	HILLIA CHALIBONENA	http://dspace.unza.zm/handle/123456789/5547	7
	KABWE KENNEDY	http://dspace.unza.zm/handle/123456789/2185	7
	JASON NGOMA	http://dspace.unza.zm/handle/123456789/6780	8
	CHILALA HABEENZU	http://dspace.unza.zm/handle/123456789/4854	6
Mines	SAMUEL KANGWA	http://dspace.unza.zm/handle/123456789/6852	5
	MICHEAL KATONGO PHIRI	http://dspace.unza.zm/handle/123456789/6877	8
	WEBBY BANDA	http://dspace.unza.zm/handle/123456789/4935	8
	CECIL DULU NUNDWE	http://dspace.unza.zm/handle/123456789/5068	6

	EDWARD CHISAKULO	http://dspace.unza.zm/handle/123456789/6335	7
Medicine	DR. CHIMOZI TEMBO	http://dspace.unza.zm/handle/123456789/7348	7
	ZIMBA CHRISTOPHER	http://dspace.unza.zm/handle/123456789/7347	6
	FRED MAATE	http://dspace.unza.zm/handle/123456789/7346	7
	FRANK CHANGWE	http://dspace.unza.zm/handle/123456789/7345	7
	DR MALAO MULEMWA BRIAN	http://dspace.unza.zm/handle/123456789/7344	6
Veterinary Medicine	LAMSON MUGALA	http://dspace.unza.zm/handle/123456789/4696	7
	HARVEY KAKOMA KAMBOYI	http://dspace.unza.zm/handle/123456789/4475	6
	CHISONI MUMBA	http://dspace.unza.zm/handle/123456789/5760	6
	CHISONI MUMBA	http://dspace.unza.zm/handle/123456789/1804	6
	JOSEPH SICHONE	http://dspace.unza.zm/handle/123456789/6720	6
Natural Sciences	MWANSA MALAMA	http://dspace.unza.zm/handle/123456789/7244	6
	SYDNEY CHIPILI	http://dspace.unza.zm/handle/123456789/7553	6
	SABBSON PHIRI	http://dspace.unza.zm/handle/123456789/7262	6
	ANNIE SWALI	http://dspace.unza.zm/handle/123456789/7645	7
	ALINANI SIMUKANGA	http://dspace.unza.zm/handle/123456789/6413	5
Graduate School of Business	PATRICK BANDA	http://dspace.unza.zm/handle/123456789/6308	6
	MPOLI MGODE	http://dspace.unza.zm/handle/123456789/6504	6
	YEDWA SONGISO MKALIPI	http://dspace.unza.zm/handle/123456789/7338	6
	NORBERT JAY MARBIN TEMBO	http://dspace.unza.zm/handle/123456789/7173	2

	CHIKUMBE SANKWA	http://dspace.unza.zm/handle/123456789/4835	7
Agricultural Sciences	HENDRIX M. CHALWE	http://dspace.unza.zm/handle/123456789/7144	6
	SIABUSU LARGEWELL	http://dspace.unza.zm/handle/123456789/7143	5
	CHISECHE MWANZA BANDA	http://dspace.unza.zm/handle/123456789/7048	6
	MAVIS .C. MUPETA	http://dspace.unza.zm/handle/123456789/6771	5
	ZOMBE KAPATA NALUPYA	http://dspace.unza.zm/handle/123456789/7843	6

Table 7. A.2: BLEU scores for Ground Truth Value and Solution Value

Record ID	Ground Truth Value	Solution Value	BLEU Score
123456789_4808	SUMBWANYAMBE	SUMBWANYAMBE	1
123456789_4876	M. C. M. BWALYA	M. C. M. BWALYA	1
123456789_5562	BANDA NZILA	ZALI IAN	0
123456789_4844	MUTALE W. CHANDA= RADHE KRISHNA	MUTALE W. CHANDA=RADHE KRISHNA	1
123456789_3767	C.M. NAMAFE	C .M.	0
123456789_4915	R. N. LIKWA= J. BANDA	R.N. LIKWA=J. BANDA	0.5841005873
123456789_4263	R JOSEPH NG'NDU	JOSEPH NG‟NDU	0.3032653299
123456789_3854	HENRY. J. MSANGO	H	0
123456789_2716	NULL	KENNY MAKUNGU	0
123456789_3651	I. A. NYAMBE	I. A. NYAMBE	1
123456789_3408	VICTOR SHIKAPUTO	VICTOR SHIKAPUTO=NUM	0.6666666667
123456789_4106	FIDELIS MUZYAMBA	FIDELIS MUZYAMBA	1

123456789_2971	S. WA SOMWE= B. AMADI	S. WA SOMWE=B. AMADI	1
123456789_3031	M. LEMBA	M Y=M. LEMBA	0.5
123456789_3032	WANGA W. CHAKANIKA	W. CHAKANIKA=A NOLT=L. H. MOONGA=E MMY MBOZI=NOAH K. SICHULA	0.1538461538
123456789_3052	A. AKAKANDELWA	A. AKAKANDELWA	1
123456789_3430	D.M. LUNGU	D.M. LUNGU	1
123456789_3011	DENNIS BANDA	BANDA DENNIS	1
123456789_4135	CHISHIMBA	MB	0
123456789_3650	WANGA W. CHAKANIKA	WANGA W. CHAKANIKA=PATRICK S. NGOMA=ANOLT L. H. MOONGA=EMMY MBOZI=NOAH K. SICHULA=ROICK CHONGO	0.1764705882
123456789_4109	R. TEMBO= B. NSEMUKILA	R. TEMBO=B. NSEMUKILA=TIM E	0.6666666667
123456789_4132	P.O.Y. NKUNIKA= E. T. MWASE	E. T. MWASE	0.513417119
123456789_3087	J. ANITHA MENON, PHD	J. ANITHA MENON	0.4776875404
123456789_3748	ELIJAH BWALYA MUTAMBANSHIKU	ELIJAH BWALYA	0.6065306597
123456789_3747	V. SESHAMANI	V. SE SHAMANI	0.3333333333
123456789_3770	ELIJAH BWALYA MUTAMBANSHIKU	ELIJAH BWALYA	0.6065306597
123456789_3753	CHARLES MICHELO=BONIFACE NAMANGALA	CHARLES MICHELO	0.3678794412
123456789_3615	CHARLES C.MICHELO= JOHN M.MILLER	CHARLES C.MICHELO=JOHN M.MILLER	1
123456789_4217	S. HATWAAMBO = M. TABAKAMULAMU	S. HATWAAMBO=M. TABAKAMULAMU=S. HATWAAMBO=M. TABAKAMULAMU	0.5
123456789_4218	WANGA, W. CHAKANIKA	WANGA=W. CHAKANIKA	0.6666666667
123456789_3722	S KAPAMBWE	WANGA W. CHAKANIKA=STANELY MPOTELA=ANOCK SAISHI=CHARLES	0

		HAKOMA=ROICK CHONGO	
123456789_3744	L. M. IMASIKU	IMASIKU=L. M.	1
123456789_4235	JOHN KINNEAR= BELLINGTON VWALIKA	JOHN KINNEAR=BELLINGTON VWALIKA	1
123456789_3764	KENNY MAKUNGU	KENNY MAKUNGU	1
123456789_3751	MAUREEN CHISEMBELE	MAUREEN CHISEMBELE=YUSUF AHMED	0.5
123456789_3774	KENNY MAKUNGU	KENNY MAKUNGU	1
123456789_3723	BEATRICE MATAFWALI	BEATRICE MATAFWALI	1
123456789_3724	D. NDHLOVU	D. NDHLOVU=GU	0.6666666667
123456789_3792	C.M. NAMAFE	PROF C.M. NAMAFE	0.6666666667
123456789_3842	F. M. GOMA= K CHOONGO= L PRASHAR	F. M. GOMA=K CHOO NGO=L PRASHAR=DEPAR TMENT	0.6
123456789_3622	MWAPE LONIA = SOKA NYIREND	MWAPE LONIA	0.3678794412
123456789_4246	J.M ZULU = M. MAKASA	J.M ZULU=M. MAKASA	1
123456789_4242	TREVOR KAILE	TREVOR KAILE=HAMAKWA MANTINA=ERIC NJUNJU=CLIVE SHIFF	0.25
123456789_4166	S.O.C MWABA	MWABA S.O.C.	0.5
123456789_3377	SELESTINE NZALA = CALLIE SCOTT	SELESTI NE NZALA=CALLIE SCOTT	0.6
123456789_3145	BELLINGTON VWALIKA	BELLINGTON VWALIKA	1
123456789_3129	BENJAMIN ANE = AGGREY MWEEM = SHABIR LAKHI	BENJAMIN ANEWS=AGGREY MWEEMBA=SHABIR LAK=BENJAMIN ANEWS=AGGREY MWEEMBA=SHABIR LAK	0.25
123456789_3846	W. W. CHAKANIKA	W. W. CHAKANIKA	1
123456789_3715	DAVIES M. LUNGU= MICK S. MWALA=	DAVIES M. LUNGU=MICK S. MWALA=OBED I. LUNGU	1

	OBED I. LUNGU		
123456789_3149	MUHAU TABAKAMULAMU	MUHAU TABAKAMULAMU=G UIDANCE	0.5
123456789_4152	ANTHONY MUSONDA	ANTHONY MUSONDA	1
123456789_3157	CELESTINE NZALA= OLIVER MWEEMBA	CELESTINE NZALA	0.3678794412
123456789_3133	CHRISPIN MPHUKA	CHRISPIN MPHUKA	1
123456789_3132	JASON MWANZA	JASON MWANZA	1
123456789_3158	CLIVE DILLONMALONE	CLIVE	0.3678794412
123456789_3654	C. SHEPANDE	VICTOR SHITUMBANUMA	0
123456789_4183	S. KASONDE-NGâ€ŸANDU = O.C. CHAKULIMBA	S. KASONDE=O.C. CHAKULIMBA	0.75
123456789_3239	J. C. MOMBA, AND ALSO M. C. BWALYA	J. C. MOMBA=M. C. BWALYA=W HO	0.625
123456789_5633	D. BANDA= G. MULEYA	D. BANDA=G. MULEYA	1
123456789_4133	KENNY MAKUNGU	KENN	0
123456789_4881	CHARLES M. NAMAFE	CHARLES M. NAMAFE	1
123456789_4471	MUSSO MUNYEME= BERNARD HANGâ€™OMBE	MUSSO MUNYEME=BERNARD HANG’OMBE	0.75
123456789_3152	F.D YAMBA=P. CHISALE	F.D YAMBA=P. CHISALE	1
123456789_4403	CHILESHE LUKWESA= GEOFFREY KWENDA= J. MWANSA	CHILESHE LUKWE=GEOFFREY KWENDA=J. MWANSA	0.8333333333
123456789_3406	ANITHA J. MENON= RAVI PAUL	ANITHA J. MENON=RAVI P AUL	0.6666666667
123456789_4874	G. MASAITI	G. MASAITI	1
123456789_3548	M.S. MWALA = M.MATAA = B.	M. S. MWALA=M. MATAA=B. DAS	0.4285714286

	DAS		
123456789_3533	G. N. SUMBWA	G. N. SUMBWA	1
123456789_4227	AIANA G. BUS	ADRIANA G. BUS=BEATRICE=MATAFWALI=DR KASONDE-NG'ANDU	0.2857142857
123456789_3418	E. PHIRI	E. PHIRI	1
123456789_4219	RAVI PAU= GIL BLACKWOOD =	RAVI PAUL=GIL BLACKWOOD=DALILA ZACHARY=HO=RAVI PAUL=GIL BLACKWOOD=DALILA ZACHARY=HO	0.2142857143
123456789_4196	WANGA, W. CHAKANIKA	WANGA=W. CHAKANIKA SIGNATURE	0.5
123456789_3267	HENRY J. MSANGO	LECTUR=HENRY J. MSANGO	0.75
123456789_3546	DALILA ZACHARY = BEN ANEWS	DALILA ZACHARY	0.3678794412
123456789_3656	CHARLES MICHELO= KNUT FYLKESNES	CHARLES MICHELO=KNUT FYLKESNES=ENCOURAG EMENT	0.6666666667
123456789_5612	MWEEMBA	MWEEMBA	1
123456789_3802	D NDHLOVU	D. NDHLOVU	0.5