

# **LIS 4014 RESEARCH PROJECT REPORT**

## **Investigating the effectiveness of integrating Controlled Subject Vocabulary Sets in The University of Zambia Institutional Repository**

### **By:**

Chipangila Bertha	13000438
Liswaniso Eric	15058590
Mawila Andrew	15014576
Mwanza Philomena	15018148
Nawila Daisy	15019551

Supervisor: Dr. Lighton Phiri

Department of Library and Information Science  
University of Zambia

November, 2019

## **Abstract**

The University of Zambia Institutional Repository (IR) has seen an improvement in the amount of digital objects being deposited therein. However, the system interface seems not to utilize subject-specific controlled vocabulary sets, leading to difficulties on the part of contributors in depositing their materials in the right location. The lack of controlled vocabulary sets also leads to ineffective search and browsing of the IR by other users of the IR. Hence the need to integrate subject controlled vocabulary sets into the UNZA IR cannot be over emphasized.

## **Acknowledgements**

Foremost, we would like to express our sincere gratitude to our supervisor Dr. Lighton Phiri for the continuous support of our LIS 4014 research project, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us immensely in our research and writing of this research paper. We could not have imagined having a better supervisor and mentor for our study.

Further, we would like to thank the LIS 5310 class, Department of Library and Information Science lecturers, UNZA library staff and our class mates for their encouragement, insightful comments, and hard questions.

Last but not least we would like to thank our families and God almighty. We could not have made it without you.

## **Dedication**

After all the hard works and effort, we would like to dedicate this study to the following. To our families, we would like to say thank you for all the love and the support that you gave us. Thank you for allowing us to stay up late just to finish this research. To our friends, we appreciate the words of wisdom that you gave. You guys have been our best cheerleaders. To our supervisor, Dr. Lighton Phiri, for educating and supporting us in doing our works, for correcting our research paper and also for guiding and motivating us, we are truly grateful. And lastly, we dedicate this piece of work to God Almighty, our source of Inspiration, Wisdom and Knowledge. We are grateful God for giving us enough strength for this kind of research. We thank you all with our hearts!

# Table of Contents

Acknowledgements	3
Dedication	4
List of Tables	7
List of Abbreviations	8
Glossary	10
CHAPTER 1	11
1. Introduction	11
1.1 Background	13
1.2 Problem Statement	14
1.3 Objectives	14
1.3.1 General Objective	14
1.3.2 Specific Objectives	14
2.3 Metadata	19
2.3.1 Descriptive Metadata in Search and Browsing	21
2.3.2 Metadata Standards	21
2.3.3 Dublin Core	22
2.4 Integration of Controlled Vocabulary Sets in DSpace	23
2.5 Conclusion	25
CHAPTER 3	26
3. Methodology	26
3.2 Study Context	26
3.3 Research Designs	26
3.3.1 Tagging of Digital Objects (how oai-pmh was used to harvested)	26
3.3.2 Subject Controlled Vocabulary Sets at the UNZA	27
3.4 Data Analysis	28
3.5 Limitations	28
3.6 Anticipated Outcomes	28
3.7 Ethics	29
Chapter 4	30
4.1 Results	30
4.2 Current Use of Controlled Vocabulary Sets	30

4.3	Familiarity with Controlled Vocabulary Sets	31
4.4	User Satisfaction	34
4.5.	Feasibility and Usability of the IR	36
Chapter 5		38
5.	Discussion	38
5.1	Use of Controlled Vocabularies	38
Chapter 6		41
6.	Conclusion	41
6.1	Summary of Findings	41
6.2	Recommendations	41
References		42
Appendices		44
Appendix 1: Structured interview guides		44

## List of Figures

Figure 1: DSpace Screenshot.....	<b>Error! Bookmark not defined.</b>
Figure 2: Mean SUS Scores.....	<b>Error! Bookmark not defined.</b>
Figure 3: DSpace Screenshot 2.....	<b>Error! Bookmark not defined.</b>
Figure 4: SUS A Summary Statistics .....	<b>Error! Bookmark not defined.</b>
Figure 5: SUS B Summary Statistics .....	<b>Error! Bookmark not defined.</b>

## List of Tables

Table 1: Dublin Core Elements .....	13
Table 2: Faculty Interviewed .....	33
Table 3: Interview Results Summary .....	34



## List of Abbreviations

DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
EAD	Encoded Archival Description
HILs	Higher Institutions of Learning
IR	Institutional Repository
LCSH	Library of Congress Subject Heading
MADS	Metadata Authority Description Schema
MARC21	Machine Readable Catalogue
MODS	Metadata Object Description Schema
MeSH	Medical Subject Heading
MIT	Massachusetts Institute of Technology
NISO	National Information Standards Organization
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative-Protocol for Metadata Harvesting
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization Systems
SPSS	Statistical Package for Social Sciences
TEI	Text Encoding Initiative
UNZA	University Of Zambia
XML	Extensible Markup Language

## **Glossary**

### **C**

#### **Controlled Vocabulary**

A list of carefully selected words used to tag units of information

### **D**

#### **DSpace**

An open source repository software package typically used for creating open access to repositories for scholarly and published digital content

#### **Dublin Core**

A metadata standard used for storing information about object title, creator, or its creation date

### **I**

#### **Institutional Repository**

An archive for collecting, preserving and disseminating digital materials of intellectual output of an institution

### **M**

#### **Metadata**

A set of data that describes and gives information about other data

### **S**

#### **Self-archiving**

The act of the authors depositing a free copy of an electronic document online in order to provide open access to it

# CHAPTER 1

## 1. Introduction

Institutional Repositories (IRs) have been an important part of Higher Educational Institutions (HEIs). According to (Foster and Gibbons, 2004), an Institutional Repository is defined as “an electronic system that captures preserves and provides access to the digital work products of a community”. They have served as information storage banks where an institution deposits its noteworthy research papers and projects to both those inside and outside the institution. There are various software tools used for repositories, of which some of the most common are Greenstone Digital Library Software, GNU EPrints Archiving Software and DSpace. In our study we will concentrate on DSpace (Tansley et al, 2003).

The University of Zambia has for the longest time deposited most of its intellectual output into its repository hosted on DSpace and continues to do so. However, to navigate these IRs, there has been a vital need to properly use digital metadata elements such as controlled vocabularies.

According to (NISO, 2014), metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. There are three types of metadata namely: Descriptive, Administrative and Structural. Administrative metadata is used for managing and preserving objects in repository while the structural is used for storage of objects in the repository and for presentation. Descriptive metadata is used for discovery of objects.

A set of fundamental metadata elements have been devised from 1995, by an international group headed by Stuart Weibel and these elements have been given the term “Dublin Core” which comes from the home of OCLC, Dublin, Ohio. The Dublin Core has fifteen central elements and they are depicted in table 1 below.

<b>ELEMENT</b>	<b>DESCRIPTION</b>
Title	Name of publication.
Creator	Author; person or organization responsible for creating the publication
Subject	Topic of the publication; usually expressed as keywords or phrases that describe the content of the resource.
Publisher	Entity responsible for making the resource available in its present form.
Format	The file format, dimensions, or physical medium of the resource.
Description	Textual description of the content of the resource.
Date	Date associated with the creation or availability of the resource
Type	Category of the resource
Contributor	Person or organization not specified in the creator element but has made significant intellectual contributions to the resource
Rights	Rights management statement, an identifier that links to a rights management statement , or an identifier that links to a service providing information about rights management for the resource
Identifier	String or number used to uniquely identify the resource, such as Universal Resource Locators(URLs)
Source	Information about second resource from which the present resource is derived

Language	The language of the intellectual content of the resource
Relation	Identifier of a second resource and its relationship to the present resource.
Coverage	Spatial locations and temporal duration characteristic of the resource.

**Table 1: Dublin Core Elements**

These elements have been essential in giving structure to IRs and thus providing ease to searching and browsing of IRs.

This study, then, aims at highlighting the problems caused in searching and browsing of the UNZA IR by a lack of use of controlled vocabularies, thereby, coming up with a strategy to integrate the use of missing metadata elements, more specifically, descriptive metadata or controlled vocabulary sets into the DSpace submission workflow.

### **1.1 Background**

The purpose of establishing the IR was to increase the accessibility of information resources and reduce on the dependence of print collection. It was noted that not enough information was being contributed by the university to the global space of information; therefore, the need of an institutional repository was evident. The University of Zambia Library was determined to share the university’s treasure of information, hence the establishment of the UNZA IR.

Since the establishment of the UNZA IR, a number of publications created by the members of UNZA and external contributors have been added to DSpace. Unfortunately, there have been inconsistencies in the submission of intellectual output to the DSpace. Therefore, it is the intention of this research is to fill in the gaps by highlighting the need to incorporate controlled vocabulary sets in the current DSpace submission workflow and investigative assess the effect of the approach.

The Institutional Repository (IR) was established in 2012, meaning from 1966 to about 2011 the University had no repository and had been working with special collections where the university collocated the published materials including theses and dissertations.

For example, when staff is entering the published works in the IR they lose concentration and put the wrong metadata, which is by either a wrong spelling, their name in place of the author's name. Lack of funding, for an institutional repository to be functional the IR it needs funding to support costs and expenses that come with running it, such as upgrades made to the system, workers payments and also machines for converting physical books to digital files. It is difficult to carry out these functions effectively with little or no funding.

The other issue raised was low published materials as an institution; it is greatly affecting how much content is deposited in the repository. The university faculties as well as students need to be publishing their works, when the output of published materials is low then the repository won't have much information for the world to search for. Lastly, before an item is deposited in the repository it has to be verified so as to prove its authenticity and quality. Not every published resource is submitted therefore poorly done work is rejected until it meets the standard of materials needed to be deposited.

## **1.2 Problem Statement**

The UNZA IR has seen an improvement in the amount of digital objects being deposited therein. However, the system interface seems to use controlled vocabularies only that they appear not to be subject specific. In a preliminary interview session conducted with the personnel in charge of the Institutional Repository said that there have been a number of challenges that affected the running and bringing the Repository at full operation. These are some of the problems that she explained. Human error has contributed to the slow development of UNZA IR and Self-archiving is usually slow and difficult without the use of subject controlled vocabularies as the people responsible for uploading the content in the DSpace find it hard to come up with the subjects that suits their field. Hence the need to integrate controlled vocabulary sets into the UNZA IR cannot be over emphasized

## **1.3 Objectives**

### **1.3.1 General Objective**

To investigate the effectiveness of using subject controlled vocabulary sets in the UNZA IR.

### **1.3.2 Specific Objectives**

1. To understand how digital objects are tagged
2. To find out which subject vocabulary sets different disciplines use at UNZA
3. To find out if the use of subject controlled vocabulary sets speeds up the ingestion process.
4. To assess the usability and usefulness of subject controlled vocabulary sets in the UNZA IR.

### **1.3.3 Research Questions**

- 1 How are digital objects tagged in the UNZA IR?
2. What subject controlled vocabulary sets are used by various disciplines at UNZA?
3. How does subject controlled vocabulary sets speeds up the ingestion process?
4. How useful are subject controlled vocabulary sets in the UNZA IR.

## **1.4 Definitions**

Institutional repository is a set of services that the university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members (Lynch, 2003).Further, Crow (2002) defines institutional repository as a digital archives of the intellectual product created by the faculty, research staff and students of an institution which are accessible to end-users both within and outside of the institution with few if any barriers to access. An institutional repository deposits are carried out to maximize visibility and accessibility of comprehensive local research which are beneficial to both the researcher and the researcher's institution.

DSpace is an open source repository software package typically used for creating open access repositories for scholarly and published digital content. DSpace repository software is a digital archives system, tool and platform focused on long term storage, access, collecting, indexing and distributing digital items. Therefore, searching is another mechanism of discovery in DSpace .Normally a user's expectation from a search is quite high ,so the goal of DSpace is to supply many search features as possible. (Smith *et al.*, 2003)

Controlled vocabulary is a list of terms and term relationship designed to collect similar information, assist content authors in consistently tagging content and enable users to find the

information they need by translating their language into the language of the information source (Leise, 2005). It is a set of selected terms used for assigning subjects in order to ensure consistency in categorizing the information to make it easier for the user find the desired information.

Metadata is simply defined as “data about data”. It is basically a set of data that describes or gives information about other data. It is a structured information that describes, explains, locates and make it easier to retrieve and use or manage an information resource (NISO, 2004).

Self-archiving is the act of the authors depositing a free copy of an electronic document online in order to provide open access to it. They usually refer to peer-reviewed research journals, conference articles, theses and book chapters deposited in the author’s own institutional repository for the purpose of maximizing its accessibility, usage and citation impact (Harnad, 2001).



## CHAPTER 2

### 2. BACKGROUND AND RELATED WORK

#### 2.1 Introduction

This chapter is a critical review of the literature relevant to this research. It will discuss in detail the integration of controlled vocabulary sets in an IR, different types of metadata with a focus on descriptive metadata as well as explain the Dublin Core metadata elements. It will conclude after a discussion on how DSpace facilitates the integration of controlled vocabulary sets in IRs.

#### 2.2 The Integration of Controlled Vocabulary Sets

A number of studies have been conducted about the integration of controlled vocabulary sets and how they facilitate improved search and browsing of Digital Libraries. Some have argued in favour of and others against the use of controlled vocabulary sets.

Controlled vocabularies are important in assisting users in their search for bibliographic information. It is assumed that by controlling vocabulary it is possible to systematically correct some of the sloppiness in language that causes the problem in retrieval (Svenonius, 1979).

The use of controlled vocabulary systems is part of a long practice of bibliographic description in the library world. Digital libraries have adopted the fundamental principles of authority control as well as many tools from the print environment. A controlled vocabulary is defined as a list or database of subject terms in which all terms or phrases representing a concept are brought together. Often one of the terms or phrases is selected as the preferred term or authorized phrase to be used in metadata records in the retrieval tool (Taylor and Joudrey, 2008).

In addition to subject terms, controlled vocabularies can include names of persons, bodies, places, objects, event and format. The term covers a wide range of tools for organizing information retrieval, but to a smaller extent, a controlled vocabulary contains a restricted list of terms. If a metadata element is designed as controlled, only terms from the selected list may be used for entry in metadata records (Hedden, 2008). The use of controlled vocabulary helps the users of the institutional repository in the searching of information they need as well as to avoid retrieval of irrelevant information

They are various types of controlled vocabularies used in the IR namely Library of Congress Subject Headings (LCSH) which is maintained by the Library of Congress to provide more precise subject description. Most libraries use the LCSH because it has a hierarchical structure with broader, narrower, and related terms specified in the headings which helps in the clarifying of the number of varied items into related subjects (Rolla, 2009).

The control aspect of controlled vocabulary comes from the fact that it determines which terms will be used to describe resources in the certain subject and that they are rules that should govern how the terms should be used as well as which terms to be added. In other words, controlled vocabulary ensures that there is consistency, accuracy and control in the search and retrieval of information (Hedden, 2008).

A study by Gross and Taylor found that more than a third of records retrieved by keyword searches would be lost without subject headings. A review of the literature since then shows that numerous studies, in various disciplines, have found that a quarter to a third of records returned in a keyword search would be lost without controlled vocabulary. Other writers, though, have continued to suggest that controlled vocabulary be discontinued (Gross, Taylor, & Joudrey, 2015).

An example of controlled vocabulary is the MeSH vocabulary which was introduced in 1960 by the United States National Library of Medicine to organize the medical literature and facilitate retrieval in the special field (The National Library of Medicine, 2013). MeSH is a standardized vocabulary of approximately 20000 terms that describe the biomedical concepts covered in the Medline database. It consists of a set of terms that are arranged in both alphabetical and hierarchical structure (Kollef, 2006).

The usefulness of MeSH terms in biomedical searching is especially important because of the extreme popularity of the PubMed central, the publicly accessible version of MEDLINE on the web.

The Georgia State University uses the MeSH vocabulary in the search of medical literature. For instance, if you want to search cancer of the jaw finding the MeSH term will help narrow down the results to further subdivisions like Jaw Neoplasms into additional categories like palatal neoplasms.

This study is relevant to our study as it talks about the importance of using controlled vocabularies as some information may be lost without them. Therefore, the study will cover the aspect of how effectiveness integrating of vocabularies in the UNZA institutional repository.

### **2.3 Metadata**

The term metadata has been increasingly adopted and co-opted by more diverse audiences; the definition of what constitutes metadata has grown in scope to include almost anything that describes anything else. Metadata are literally or technically ‘data about data’ or information about information or information that makes data useful. Moreover metadata as data whose primary purpose is to describe, define and or annotate other data that accompanies it. The structured data of Metadata describes the characteristics of a resource (Abidillah, 2013).

A metadata record consists of a number of predefined elements representing specific attributes of a resource, and each element can have one or more values. It is an extensive and expanding subject that is prevalent in many environments. They provide information on such aspects as the ‘who, what, where and when’ of data and can be considered from the perspective of both the data producer and the data consumer. In terms of search, metadata is very useful key for search engine to recognize as the guide about what information should be provided to the users and it also determines the level of success of a search. The most efficient way to make search work better is to bring some metadata to bear on the problem because metadata are used for searching and scientific papers usually have certain pieces of metadata (usually assigned by authors) describing the topics.

There are five main types of metadata namely: Administrative, Technical, Structural, Preservation and Descriptive metadata. Administrative metadata is metadata used in managing digital objects and collections and information resources which further aids to know when and how a digital object is created, file type and other technical information about who can access it.

Administrative metadata is helpful for both short term and long term managing of digital collections according to logically defined needs to secure its integrity for instance rights of management ,access ,control and use requirements ,acquisition information ,location information and selection criteria for digitization.

Technical metadata on the other hand is metadata related to how a system functions or metadata behaves. Structural metadata enables navigation and presentation of electronic resources. It documents how the components of an item are organized. For example: how pages are ordered to form chapters of a book.

Metadata related to the preservation management of collections and information resources is known as preservation metadata for instance file type and extension , software needed to open digital files ,actions taken to preserve physical materials and digital files ,documentation of any changes occurring during digitization or preservation.. Metadata related to how a system functions or metadata behaves is known as Technical metadata for instance software's documentation, technical digitization tracking of system response which includes the time, authentication and security data. Lastly, descriptive metadata is used to identify and describe collections and related information resources. (ISQ, 2010)

In a study which was conducted in spring 2014, authors from the University of Missouri conducted a nationwide survey on metadata practices among United States –based OpenDOAR repository examining the repository systems of the institution. In the study, they concluded that the usefulness of metadata practices is dependent on many factors including the system functionality, the encoding of metadata for the machine manipulation and the quality of the metadata. In the study they gathered information on systems used, metadata encoding schemes and elements that impact the quality of metadata which included the level of staff creating it. In the study, the creation of metadata for research and repository content is essential part of scholarly communication process and is necessary for a long –term access and preservation of our digital heritage. Metadata choices and particles affect the find ability of resources in the online environment and these choices, influenced by the content itself also reflects the institution itself, stakeholders and users of specific repositories. (Heather et al, 2015).

In a nutshell, metadata can be identified as the foundation of all information retrieval. It is the key to guarantee that resources will stay alive and continue to be accessible into the future. Without adequate metadata, one is not able to locate information sources any more. The non-retrievable documents then become traceless, forgotten or most likely deleted. As a result, this unplanned loss of information may have significant and costly penalty for a society. (Rahman et al, 2011).Our study focuses on descriptive metadata.

### **2.3.1 Descriptive Metadata in Search and Browsing**

Descriptive metadata describes a resource for the purpose such as discovery and identification. It includes elements such as author, title and keywords (NISO, 2004). Descriptive metadata further includes the bibliographic metadata, which provide a bibliographic description of the publication and are used for retrieval purposes. They also include the structural metadata providing information about the relations between parts of publications such as serial title, issues and articles that are used for electronic collections browsing. Part of the metadata may have been created by the author and by the publisher. The librarians check and improve the metadata by creating links with authority files for the main access points or by organizing the items into the electronic collection structure. The metadata is used directly by the users to select the electronic publications in which they are interested (Lupovic & Masanes, 2000).

This shows that descriptive metadata is critical in facilitating search and browsing. Thus the lack of use of descriptive metadata in an IR would result in difficulties in search and retrieval of digital objects.

### **2.3.2 Metadata Standards**

There are a number of metadata schemes used in descriptive metadata: MARC21 (Machine Readable Catalogue), it is a set of codes and content designators defined for encoding machine readable records. They are standards for the representation and communication of bibliographic as well as related information in Machine Readable form

MODS (Metadata Object Description Schema), it is an XML schema that was created to encode descriptive metadata on digital objects.

MADS (Metadata Authority Description Schema), is an XML schema for an authority element set that may be used to provide metadata on agents, events and terms. It serves as a companion to the Metadata Object Description schema to provide metadata about authoritative entities used in MODS descriptive

EAD (Encoded Archival Description) is a standard for encoding descriptive information regarding archival records. It allows users to locate primary sources that are geographically remote.

TEI (Text Encoding Initiative), it is a consortium which collectively develops and maintains a standard for the representation of texts in digital form DC (Dublin Core) (Hayes, 2008).

### **2.3.3 Dublin Core**

The Dublin Core is the most widely used metadata standard. General Metadata standards, like the Dublin Core allows storing information about object title, creator, or its creation date. The Dublin Core metadata is a set of the 15 elements designed to foster consensus across disciplines for the discovery oriented description of diverse resources in an electronic environment. It can be applied to any electronic resources from graphics to sound (Chmielowski, 2007). The elements include the Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

The Dublin Core Element Set (DC) was developed in 1995 as a means in which to improve indexing of search engines by embedding metadata elements into web pages or encoding through the use of XML. This metadata standard was created to increase interoperability of metadata records, by bridging the differences of the existing objects descriptions. It is a common denominator of existing metadata standards, it has only 15 optional and repetitive elements that are very generic and clear in context, and they represent semantic crosswalks among metadata standards in different disciplines. Dublin Core (Dublin Core Metadata Initiative, 2011a) is an easy to learn and use schema that is a basic default metadata template in many digital content management systems. The goals of Dublin Core are simplicity and ease of use, commonly understood semantics, international scope, and extensibility. It was created to be intentionally “generic,” allowing user communities to define content standards and the use of controlled vocabularies that fit specific needs. The interoperability of Dublin Core metadata fields makes it easy to share data and create discovery opportunities.

A project was undertaken in 2010 at the University of Salamanca with the goal of creating a digital institutional repository called GREDOS that would respond to the demands of the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH) protocol, to be able to offer to the members of the academic community and to society, in general, access to, dissemination of, and preservation of the digital material created by the institution and its members, as well as the digitalized holdings that make up the rich bibliographical and documental heritage of the University based on the Dublin Core Metadata Initiative (DCMI).

The GREDOS Institutional Repository showed that Developing metadata best practices in line with standards achieves assurance of quality in the metadata records; increased possibility of discovering the resource; increased interoperability of the GREDOS collections, increased interoperability among other digital repositories and libraries participating in the OAI (Open Access Initiative); ease in being picked up by contents providers such as DRIVER; information to users about the structure of the digital object and the visualizer necessary for accessing the digital resource; and assistance in the management of long-term preservation of digital archives. (Peñalvo et al., 2010).

#### **2.4 Integration of Controlled Vocabulary Sets in DSpace**

DSpace is designed in a way that it can support multiple controlled vocabulary sets. One study by Solomou and Koutsomitropoulos aimed to show this by integrating one type of controlled vocabulary systems that is not, by default, supported by DSpace and this is known as the Simple Knowledge Organization System (SKOS) which provides a standard way of representing knowledge organization systems using Resource Description Framework (RDF). In DSpace, SKOS is implemented through an add-on, provided by the University of Minho.

SKOS is data model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, taxonomies and other similar types of controlled vocabularies. It is a practical application of RDF and mainly enables easy publication of controlled structured vocabularies for the Semantic Web

In this study, the add-on is first re-implemented then finally applied to the University of Patra's live DSpace installation. The experiment is then specifically done by importing a real SKOS vocabulary: the thesaurus of Greek Terms. The provided add-on offers the ability to assign a different vocabulary to each community in the University's Institutional Repository.

The study revealed a few problems that arose during the experiment. The given SKOS-to-node XML transformation proved difficult due to the transformation copying with only the narrower terms described also as standalone concepts. Further, some terms in the constructed vocabulary set were either absent or repetitive (Solomou & Koutsomitropoulos, 2014).

This paper is relevant to our study because we will integrate multiple vocabularies, not just one which will help the users in the searching and browsing of information in the IR.

In a research study by Ferreira and Baptista (2005), The use of Taxonomies as a way to achieve Interoperability and improved Resource Discovery in DSpace-based Repositories developed at University of Minho in the context of Institutional Repositories, more precisely the ones based on the DSpace platform (developed jointly by the Massachusetts Institute of Technology and Hewlett-Packard) was done to create an add-on for DSpace to ensure authority control over the keywords that human cataloguers may use to describe their items of information. These keywords are used by the visitors of the repository for searching and browsing the catalogue. The keywords are organized in a taxonomy that results from the combination of several specialized thesauri, one for each community of users in Computing, Engineering, and Architecture can use. Taxonomies were described as a subject-based classification that arranges the terms in a controlled vocabulary into a hierarchy without doing anything further. Almost anything animate objects, inanimate objects, places, and events may be classified according to some taxonomic scheme. Mathematically, a taxonomy is a tree structure of classifications for a given set of objects. At the top of this structure is a single classification the root node that applies to all objects. Nodes below this root are more specific classifications that apply to subsets of the total set of classified objects. Hence, a taxonomy is a collection of terms used to describe things that are grouped together in a tree structure. We are able to identify parent-child relationships between the terms in the controlled vocabulary. During the add-on of controlled vocabulary, it was observed that in most archiving scenarios, it is natural that a certain degree of ambiguity and heterogeneity will be found in the metadata provided by different users to documents with similar content. This can also be observed in archives where items have been indexed by trained professionals. To downsize this problem, an add-on was developed for the reason that it restricts the keywords that users may employ during indexing stages of self-archiving.

Furthermore, during submission, users were asked to enter the keywords that best describe their works. With the add-on in place, users were presented with a taxonomy that displays the terms that are allowed to be used as descriptors. For each community of users that interact with the repository a different taxonomy is presented. Each of these taxonomies is rendered to the user as an expandable tree. For example, the first controlled vocabulary that was imported into the



system was the ACM Computing Classification System. This controlled vocabulary is being used by the students of the Department of Information Systems to describe their academic projects. Recent contacts with other departments also interested in publishing their students' projects have resulted in the opening of the Papadocs repository to the Civil Engineering and Architecture communities. This event conducted to the adoption of two other taxonomies appropriate to describe the items submitted by members of these communities we are now using a sub-set of the Engineering Index Thesaurus and negotiating the possibility of using the Art & Architecture. The add-on works by loading all the included taxonomies from independent XML files (stored on the server's file system) and rendering them as trees to the user. The structure of these XML files is very straightforward. We use four different elements to represent the whole structure of the taxonomies: node, which contains information about a specific term; is Composed By, a wrapper element that contains a list of child nodes; is Related to, an element that contains links to other related nodes in the taxonomy; and has Note, an element that allows the inclusion of a small descriptive note about the term. With this it was observed that the advantage of this approach is twofold. It removes the ambiguity inherent to certain concepts by accompanying them with the correct context and allows the realization of more general queries and that the introduced controlled vocabulary system is adequate to our objectives due its simplicity. Users can easily find the terms they are looking for by expanding just a few branches of the taxonomy.

This study is important to our research as it tries to bring out the concept of making sure that communities representing a particular field of study should be represented by various terms which can make it easy for resource discovery. It further outlines that through submissions done by users of DSpace should be presented with various controlled vocabularies which removes the ambiguity of certain terms but instead give them a list of possible terms to be used.

## **2.5 Conclusion**

The above chapter was the literature review which explained the different literature on the research. Themes were designed in order to make it easier to explain the literature. The first theme was the integration of controlled vocabulary sets. Another theme explained the types of metadata used in the institutional repository. Finally, it explained how the DSpace facilitates the integration of controlled vocabulary sets.

## CHAPTER 3

### 3. Methodology

This section describes the methods employed to investigate the problem at hand. It includes the research design, study site, sample size, data collection instruments, data analysis tools and results, as well as limitations encountered during the study.

#### 3.1 Research Approach

This study followed the Pragmatic philosophical worldview which does not focus on methods but emphasizes the research problem and uses all approaches available to understand the problem (Rossman & Wilson, 1985). Therefore, it applied to mixed methods research which was the exact approach that this study took.

The mixed methods approach involved both qualitative and quantitative techniques. The specific procedures, measurement instruments and analysis will be discussed in the sub sections that followed.

#### 3.2 Study Context

The research was carried out at the University of Zambia (UNZA) Great East Road Campus. The targets were faculty staff, key handlers of the IR and the few students who were randomly selected to measure the effectiveness of searching and browsing the IR after integrating subject controlled vocabulary sets in a prototype. That way, we got detailed information relevant to the study.

#### 3.3 Research Designs

##### 3.3.1 Tagging of Digital Objects (how oai-pmh was used to harvested)

The initial stage of the study required us to establish how digital objects are tagged in the UNZA IR. To do this we harvested metadata for all the Electronic Theses and Dissertations (ETDs) from the UNZA IR using the OAI-PMH (Open Archives Initiatives Protocol for Metadata Harvesting) and analysed the data using Excel and STATA. ETDs were targeted because they are cross cutting across all academic fields and hence provide a representative sample. The OAI-PMH is a metadata harvesting interoperability framework (Lagoze et al., 2002).

The OAI-PMH was used to harvest metadata from HEIs IRs, using the LibreCat Catmandu data processing toolkit. The harvesting was done using the Dublin Core metadata format—`metadataFormat=oai_dc`. In addition to the `SetSpec` field of the harvested metadata, the subject Dublin Core elements were used during the analysis stage. Resources associated with each digital object were harvested using the Open Archives Initiative Object Reuse Exchange (OAI-ORE) standard —`metadataFormat=ore`.

After this, we conducted structured interviews with two staff members from the UNZA library who work with the repository to establish what controlled vocabulary sets are used when depositing scholarly output into the IR. The two were selected using purposive sampling because they are the ones that work with the repository on a daily basis and thus are better placed to explain what controlled vocabularies, if any, are used when tagging digital objects.

Purposive sampling method is a non-random sampling technique in which particular settings, persons or events are selected deliberately in order to provide important information that cannot be obtained from other choices. It is where the researcher includes participants in the sample because they believe that they warrant inclusion. The main reason for the use of purposive sampling is that there are only a few individuals to interact with and you know these individuals. (Tongco, 2007).

Further, semi-structured interviews were used to avoid researcher bias and for easy analysis. It further allowed the researcher to focus on the specific information required for the research with questions focused on specific objectives and is often considered a more effective way of testing the respondent.

### **3.3.2 Subject Controlled Vocabulary Sets at the UNZA**

The next phase of the study involved further semi-structured interviews with key stakeholders from the various faculties at the UNZA. The 10 faculty staffs were also selected using purposive sampling for reasons outlined in Section 3.3.1. The main goal of these interviews was to establish what controlled vocabularies are commonly used in their various fields.

### **3.3.3 Effectiveness of integrating Controlled Vocabularies within IR**

The final phase of our study involved us experimentally setting up a sandbox where we installed two instances of DSpace (DSpace\_baseline and DSpace\_control) on an external server (<http://lis.unza.zm:8081/xmlui/>) and integrated some subject controlled vocabulary sets into the baseline. The selected controlled vocabularies were determined by the findings from our interviews with faculty staff. Thus their inclusion was only logical. This process also helped us assess the effectiveness of DSpace in integrating different controlled vocabulary sets and whether it would lead to improved search and browsing of the IR.

After integrating the controlled vocabularies into DSpace, experts who work within the Library were used to test the submission workflow. The aim of this was to determine whether the integrated controlled vocabularies helped speed up the ingestion process or not.

Additionally, the approach also helped us evaluate whether the integration of subject controlled vocabularies led to improved search and browsing of the IR. To do this, 50 Library and Information Science students from the University of Zambia were used. The students were exposed to both DSpace\_baseline and DSpace\_control and asked to evaluate their experience. The metrics used to evaluate the effectiveness of this approach were precision and recall. We used a within subject design because it is cheaper and requires a smaller sample size.

### **3.4 Data Analysis**

We manually analyzed the responses from the interviews to determine the most commonly used subject controlled vocabularies. This is what informed our choice of controlled vocabularies for our sandbox.

### **3.5 Limitations**

The major limitation of our study was the unavailability of related literature to the specifics of our study. It appeared that no much research has been done about the use of controlled vocabularies in IRs.

### **3.6 Anticipated Outcomes**

At the end of the study, it was anticipated that there would be a push towards self-archiving. Another anticipated outcome was that the use of controlled vocabularies was to facilitate comprehensive description of content deposited using self-archiving. It was further anticipated

that this would lead to ingested content having fewer errors since depositors of digital objects could choose subjects from a dropdown menu rather than provide open-ended text. Lastly, that content was consistently tagged with similar subjects, making it relatively easier to find related content.

### **3.7 Ethics**

Prior to this study, ethical issues were properly addressed. This included getting consent from relevant sources of information and authorities, and everyone that shall be involved in the study. Ethical clearance was sought after in order to uphold the integrity of this research, to respect and protect the confidentiality of all participants. Therefore, it was imperative that the purpose of the study, duration of the study and any benefits be clearly explained to all participants.

Safety measures were taken into consideration by making sure that the study was conducted in a peaceful and harmless manner as it is the duty of the researcher(s) to ensure that the safety of all participants is guaranteed.

Participants involved in the study were given the autonomy to answer questions asked to them or to decline if they felt uncomfortable. Therefore, no coercion was used to make anyone participate in the study; they exercised their freedom to participate or not. The identities of those that decide to participate were not revealed confidentiality was kept from the beginning to the end of the study. Care was taken in ensuring that no breach in the privacy of these participants occurs.

## Chapter 4

### 4.1 Results

This chapter reports the results that were obtained after conducting interviews with library staff, faculty staff from different faculties, and a quasi-experiment in which students from the School of Education participated.

### 4.2 Current Use of Controlled Vocabulary Sets

Interviews were conducted with two library staff, the repository manager and his assistant, in charge of the repository in order to find out how scholarly materials are ingested in the University of Zambia institutional repository and to find out if the use of subject controlled vocabulary sets is incorporated. The repository manager explained that before material is ingested into the repository, it is first catalogued then metadata for each material is copied from OPAC then pasted into the repository. The library is said to use Library of Congress Subject Headings but it has not been integrated into the University of Zambia DSpace.

From the table below we can see that some records which are supposed to be grouped under one subject are appearing separately when they are closely associated for example Breastfeeding and Breastfeeding. These subjects are closely associated and can be grouped under one subject but they are treated as separate subjects when they are all focused on one subject for the reason that one was typed in ending with a full stop which makes it look different when they are associated. Further in relation to the tables above some subjects are differentiated by using hyphens when they can be grouped under one title which will represent them all. Furthermore 1 record from the table has no subject assigned to it.

Binge Drinking	1	0.05	12.21
Bio-availability -- Zambia	1	0.05	12.26
Biology	2	0.10	12.36
Biology--Study and teaching	1	0.05	12.41
Biotechnology--Government Policy--Zambia	1	0.05	12.46
Birth Control	1	0.05	12.51
Birth control --Law and legislation --.	1	0.05	12.56
Birth control.	3	0.15	12.71
Birth control.--Zambia	1	0.05	12.76
Blasphemy- Zambia	1	0.05	12.81
Botswana	1	0.05	12.86
Bottle Feeding	1	0.05	12.91
Brain	1	0.05	12.96
Breasfeeding	1	0.05	13.01
Breast -- Examination	1	0.05	13.06
Breast Abscesses	1	0.05	13.11
Breast Cancer--Zambia	1	0.05	13.16
Breast Feeding	2	0.10	13.26
Breast Feeding.	2	0.10	13.36
Breast milk substituites--Zambia	1	0.05	13.41
Breastfeeding -- Diseases -- Zambia	1	0.05	13.46
Breastfeeding --Health aspects.	1	0.05	13.51
Breastfeeding --Health aspectsZambia	1	0.05	13.56
Breastfeeding --Zambia.	2	0.10	13.66
Breastfeeding--Immunological aspects.	1	0.05	13.71
Broilers (poultry) -- Zambia	1	0.05	13.76
Brucellosis	1	0.05	13.81
Brucellosis in cattle--Zambia	1	0.05	13.86
Budget--Zambia	1	0.05	13.91
Burden of Proof (insanity) - Zambia	1	0.05	13.96
Burns	1	0.05	14.01
Caesarean Section--informed Consent	1	0.05	14.06
Calcitonin	1	0.05	14.11
Calcium metabolism disorders	1	0.05	14.16
Cancer	1	0.05	14.21

Figure 1: DSpace Screenshot

### 4.3 Familiarity with Controlled Vocabulary Sets

Interviews were conducted with five faculty staff from the School of Education (four under the Department of Library and Information Studies and one under Adult Education), one from the School of Engineering and the other one from the School of Veterinary Medicine. The basis of the interviews was to find out whether faculty staff was familiar with controlled vocabulary sets in particular, the exact subject controlled vocabulary sets used in their field. The reason for this

was to identify which specific controlled vocabulary would need to be integrated into the sandbox.

Two tables were created to show how many participated in the interviews to summarize the results obtained from all interviews as shown below.

ITEM	CATEGORY	COUNT
Gender	Male	7
	Female	2
Designation	Lecturer III	
	Lecturer II	
	Lecturer I	1
	Senior Lecturer	1
	Associate Professor	
	Professor	
Years in Service	Less than 5 years	1
	5 to 9 years	3
	10 to 14 years	1
	15 to 19 years	1
	More than 20 years	1
Highest Academic Qualification	Master's Degree	6
	Doctoral Degree	1
Faculty/School	Agricultural Sciences	0
	Education	5
	Engineering	1
	Humanities and Social Sciences	0



	Law	0
	Veterinary Sciences	1
	Main Library	2

Table 2: Faculty Interviewed

<b>No.</b>	<b>Respondent</b>	<b>Familiarity with CVs</b>	<b>Faculty</b>	<b>Databases Used in Respondent's Field</b>	<b>CVs Associated with Database</b>	<b>Comments</b>
1	FS-1	Yes	School of Veterinary Sciences	PubMed, Science Direct, Google Scholar, Medley	MeSH	Respondent was aware of MeSH as a collection of key terms but didn't know that it was a controlled Vocabulary set until it was explained.
2	FS-2	No	School of Education	SCOPUS, ERIC, SCINAPSE, EBSCO HOST, PREQUEST	Not aware	Not familiar with the specific controlled vocabularies associated with the databases in respondents field
3	FS-3	Yes	School of Education	Academia, Zambia Library Journals, Google Scholar, UNZA IR	Not aware	Familiar with controlled vocabularies but not aware of the specific ones used in respondents field
4	FS-4	No	School of Engineering	IEEE, ELSEVIER	Not aware	Familiar with the concept of key terms but not familiar with any specific controlled vocabulary sets
5	FS-5	Yes	School of Education	Research Gate, Google Scholar, Academia	None	Familiar with controlled vocabularies but does not use specific controlled vocabularies when publishing
6	FS-6	Yes	School of Education	Academia, Mendeley, Research Gate, JSTOR, Google Scholar	SEARS list	Familiar with and uses controlled vocabularies
7	FS-7	No	School of Education	IEEE, Explorer, Research Gate	None	Highlighted that there are submission guidelines but not familiar with any specific controlled vocabularies
8	LS-10	Yes	Main Library		LCSH	
9	LS-11	Yes	Main Library		LCSH	

Table 3: Interview Results Summary

#### 4.4 User Satisfaction

To measure user satisfaction with the rate of ingestion, the study used students from the department of Library and Information Science at the University of Zambia. The students were exposed to two instances of DSpace; one integrated with subject controlled vocabularies and the other without subject controlled vocabulary sets. After their interaction with each instance, the students responded to an online questionnaire giving feedback on their experience. Their responses were then compared and analyzed to determine which system was preferable to the user. Thus the mean scores of the two SUS questionnaires are as captured in the graph below.

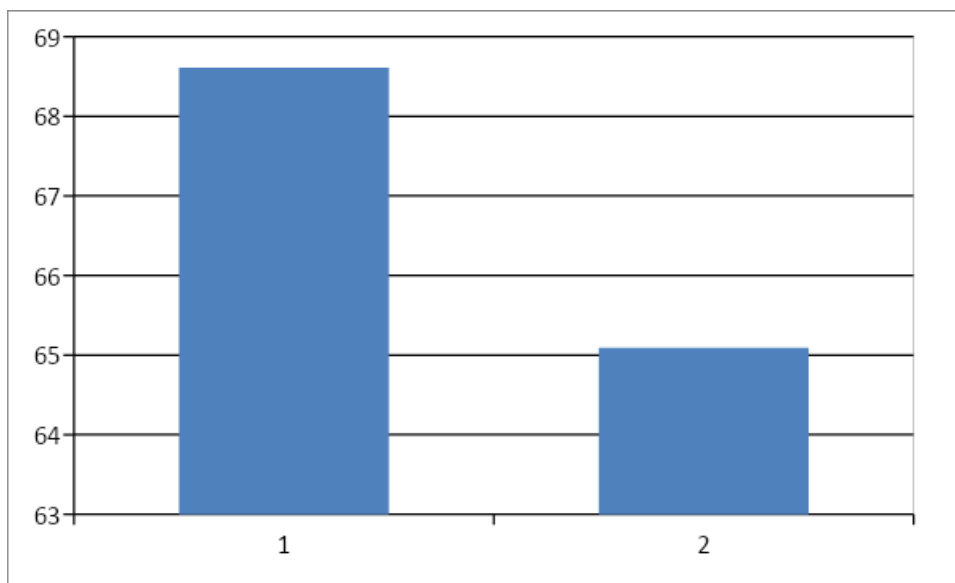


Figure 2: Mean SUS Scores

Participants in the study were asked a series of questions regarding their experience including whether or not they found the system to be cumbersome on a scale of one to five, with 1 being strongly disagree and 5 being strongly agree. The findings were analyzed and are here presented graphically.

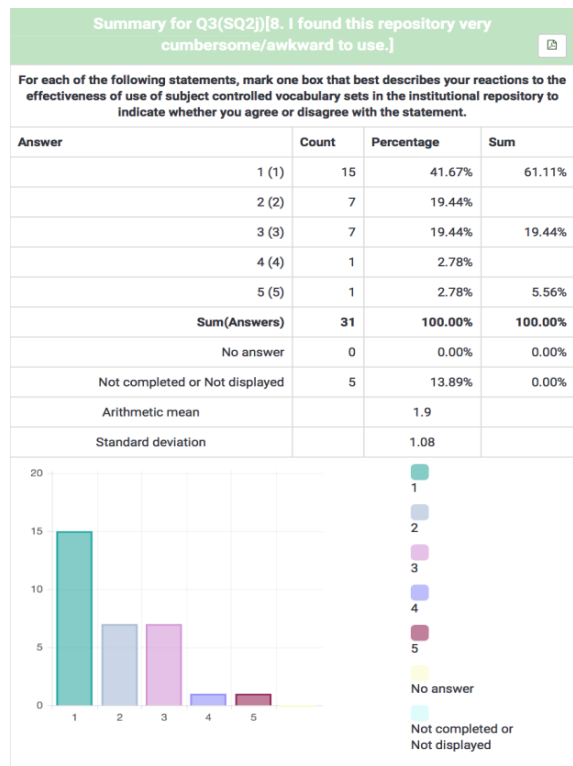
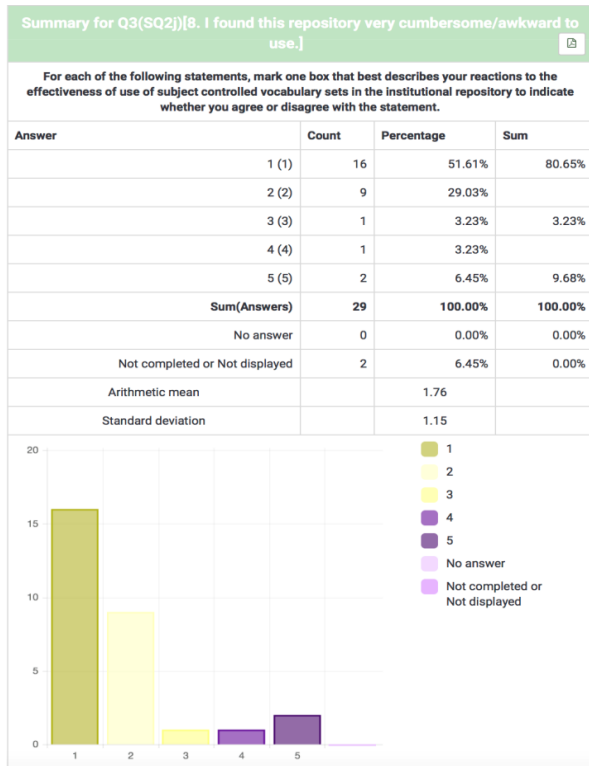


Figure 3: SUS A and SUS B

#### 4.5. Feasibility and Usability of the IR

In trying to determine the feasibility of integrating subject controlled vocabularies into DSpace, we carried out a quasi-experiment as outlined in the section above. The experiment involved us integrating LCSH into the IR and also redefining the communities and collections in the IR to reflect an idea IR. It is this instance of DSpace that the participants in the study were made to interact with. Below is a screenshot of this instance of DSpace with redefined communities that the study felt are more suited to the UNZA scenario.

## DSpace Repository

DSpace is a digital service that collects, preserves, and distributes digital material. Repositories are important tools for preserving an organization's legacy; they facilitate digital preservation and scholarly communication.

## Communities in DSpace

Select a community to browse its collections.

- [Graduate School of Business](#)
- [Graduation Speeches](#)
- [School of Agricultural Sciences](#)
- [School of Education](#)
- [School of Engineering](#)
- [School of Health Sciences](#)
- [School of Humanities and Social Sciences](#)
- [School of Law](#)
- [School of Medicine](#)
- [School of Mines](#)
- [School of Natural Sciences](#)
- [School of Nursing Sciences](#)
- [School of Public Health](#)
- [School of Veterinary Medicine](#)

### Search DSpace

[Advanced Search](#)

### Browse

All of DSpace  
[Communities & Collections](#)  
[By Issue Date](#)  
[Authors](#)  
[Titles](#)  
[Subjects](#)

### My Account

[Login](#)  
[Register](#)

### RSS Feeds




 [RSS 1.0](#)  
 [RSS 2.0](#)  
 [Atom](#)

Figure 4: DSpace Sample IR Screenshot

## Chapter 5

### 5. Discussion

This chapter is a critical review of the results in connection to the literature review. It presents a picture of the current use of and level of familiarity with controlled vocabulary sets by UNZA faculty and library staff.

#### 5.1 Use of Controlled Vocabularies

The study reviewed in the interviews conducted with the library staff that the use of subject controlled vocabulary sets in the UNZA IR are not in use because they have not been incorporated in DSpace and they had no knowledge that that could be possible. The library staff use the already catalogued metadata for each material copied from the OPAC then paste in the repository. In other instances the library staff uses to refer to the LSCH manuals to look for the titles which can suit a particular document. In addition, the library staff said the process for ingestion takes long; approximately two hours for one item to be uploaded which was cumbersome. They further suggested that the communities in the UNZA IR are not properly arranged because they were too many. Hence, integration of subject controlled vocabulary sets and arranging the communities according faculty would help the repository reduce inconsistencies that occur in coming up with subjects.

#### 5.2 Familiarity with Controlled Vocabulary Sets

The results shows that most faculty staff were not aware of subject controlled vocabulary sets that are used in their field. However, they were knowledgeable of the databases that are used when depositing their research. That way, we were able to determine the controlled vocabulary sets associated with the database by searching. From the school of veterinary, sciences the participant mentioned PubMed science direct, Google scholar and Mendeley. From the database we were able to know that MeSH is used in their field and acknowledged that they actually use it but he was not aware that it was a controlled vocabulary. Participant two from the school of Engineering stated that they use IEEE database and said that he was familiar with the concept of key terms but not aware of controlled vocabularies. From the school of education, we interviewed a lecturer who is familiar with controlled vocabularies sets and stated that they used

sears list of subject heading from the department of Library Information Science. The other four participants were not familiar with controlled vocabularies sets. Therefore, for our sandbox we integrated the LCSH to the DSpace because we were dealing with the school of education particularly department of library information science. Sears list was not possible to be integrated in DSpace hence resorted to LCSH and arranged the communities according to the faculties to make it easier for ingestion process.

### **5.3 Feasibility and Usability of the IR**

The results showed that most of the participants have an average knowledge in the use of computers. Their having such knowledge implies that they are likely to know the use of computers and results viewed of how good numbers of users are likely to have ideas and easily follow the procedures to undertake the ingestion process in DSpace. The others who had little knowledge on ICTS related courses resulted in poor rating usability of controlled vocabulary sets in the ingestion process.

On the other hand, it is important to realize that all the participants whether or not they have had ICT related courses they are not entirely disadvantaged in using the system. The participants included the Library information science students and ICT students. This is because they are likely to have experiences in using ICTS and have therefore learnt many computer skills. Hence they did not have any challenges in the ingestion process in the repository.

According to the results obtained using the SUS questionnaire, it showed that the participants had a wide range of reaction to the system. This is shown by the standard deviation of the calculated SUS scores of each participant. The standard deviation expressed how much the members of the group of participants differed from the mean value for the group. The members had different perspectives about the use of controlled vocabularies in the ingestion process and not using controlled vocabularies. This is the reason why some participants rated the system with the controlled vocabularies with the mean 68.6 while others rated the repository without controlled vocabularies with the mean 65.1. This rating entails that is in good position in terms of usability.

```
. sum v12
```

Variable	Obs	Mean	Std. Dev.	Min	Max
v12	27	68.61111	18.42779	32.5	100

.

Figure 5: SUS A Summary Statistics

```
. sum v12
```

Variable	Obs	Mean	Std. Dev.	Min	Max
v12	27	65.09259	20.24468	20	100

Figure 6: SUS B Summary Statistics

According to the comments that were given by participants, their comments showed positive attitude towards the use of controlled vocabularies in the repository and made it clear that there integration of controlled vocabularies should be implemented in the UNZA IR in order to speed the ingestion process. However, some respondents in the comment section stated that there is a need for sensitization on the use of institutional repository and thought the use of controlled vocabulary sets was cumbersome.



## **Chapter 6**

### **6. Conclusion**

#### **6.1 Summary of Findings**

The use of subject controlled vocabulary sets in the University of Zambia Institutional Repository would improve the submission work flow of digital content. The results of the research show that the use of Subject Controlled Vocabulary sets is more efficient than manually ingesting content onto the repository, from our table we clearly see that there is too much inconsistency with the data, as a result of not using the subject controlled vocabulary. Furthermore, this would encourage lectures to self-archive their own work and improve the amount of content in the repository because it's easier and faster when controlled vocabulary are integrated into the system.

#### **6.2 Recommendations**

In the light of the aforementioned discussion, the following recommendations are made:

The University of Zambia should integrate Subject Controlled Vocabularies suitable for each field of study to make the content more accessible to users of the institutional repository and improve the ingestion process of digital content.

Furthermore, the management of the IR should consider encouraging faculty staff to self-archive their work so as to lessen the burden of the library staff who are in charge of the repository and also lessen publication delays.

Lastly, the University should consider redefining the communities and the collections in the repository to the proposed format as shown in chapter 4 above, as this would help improve the ease of access to materials in the IR as well as making self-archiving easier for faculty.

## References

- Bimbe, N. B., Lungu, S., Kakana, F., Sichilima, C., Makondo, F. N. S., & Kanyengo, C. W. (2017). *IDS WORKING PAPER Volume 2017 No 483 Challenges in Reinvigorating and Upgrading DSpace-based Institutional Repositories : A University of Zambia (UNZA) Library Case Study* (Vol. 2017).
- C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner, "Open Archives Initiative-Protocol for Metadata Harvesting-v. 2.0," 2002.
- Chmielewski, J(2007) *Metadata Schema of Interactions for Multimedia Objects* ,Technical University of Gdansk
- Crow, R. (2002). *The case of institutional repositories: a SPARC position paper* (Washington, D.C.) Available at [http://www.arl.org/sparc/bm%7Edoc/ir/final release 102.pdf](http://www.arl.org/sparc/bm%7Edoc/ir/final%20release%20102.pdf)
- Gross, T., Taylor, A. G., & Joudrey, D. N. (2015). Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching. *Cataloging & Classification Quarterly*, 53(1), 1–39. <https://doi.org/10.1080/01639374.2014.917447>
- Harnad, S. (2001). The self-archiving initiative. *Nature*.410 (6832): 1024-1025. Doi:10.1038/35074210
- Hedden, H. (2008). *Controlled vocabularies, thesauri, and taxonomies. Indexer*, 26(1), 33-34
- ISQ. (2010). FE\_Dappert\_Enders\_MetadataStds\_isqv22no2. *NISO*.
- Leise, F. (2005). *Controlled vocabularies: An introduction. Indexer*, 26(3), 121-126.
- Lupovic ,C and Masanes J. (2000) *Metadata for Long-term Preservation* <https://www.kb.nl/sites/default/files/docs/preservationmetadata.pdf>
- NationalLibraryofMedicine.(2015).Retrievedfrom[http://www.nlm.nih.gov/mesh/intro\\_preface.html#pref\\_rem](http://www.nlm.nih.gov/mesh/intro_preface.html#pref_rem)
- NISO (2004) *Understanding metadata* [https://www.iter.uaf.edu>metadata\\_files](https://www.iter.uaf.edu>metadata_files)
- NISO (2014). *Understanding Metadata What Is Metadata*. Retrieved from [http://www.niso.org/apps/group\\_public/download.php/17446/Understanding Metadata.pdf](http://www.niso.org/apps/group_public/download.php/17446/Understanding_Metadata.pdf)
- NISO (2004). *Understanding metadata*. NISO Press. Retrieved from [www.niso.org](http://www.niso.org)

Peñalvo, F. J. G. *et al.* (2010) 'Qualified dublin core metadata best practices for GREDOS', *Journal of Library Metadata*, 10(1), pp. 13–36. doi: 10.1080/19386380903546976.

Purposive Sampling. (n.d.) retrived from <http://dissertation.laerd.com/purposive-sampling.php>

Rahman, A. I. M. J., Francese, E., Yilmaz, M., & Beyene, W. (2011). Metadata practices in digital libraries. *Proceedings of the International Seminar 'Vision 2021: The Role of Libraries for Building Digital Bangladesh*, (Dill).

Rolla, P. J. (2009). *User tags versus subject headings: Can user-supplied data improve subject access to library collections?* *Library Resources & Technical Services*, 53(3), 174-184.

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., & Smith, M. (2003). The DSpace institutional digital repository system: Current functionality. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2003-January*, 87–97. <https://doi.org/10.1109/JCDL.2003.1204846>

Smith, M. *et al.* (2003) 'Dspace: an open source dynamic digital repository', *D-Lib Magazine*, 9(1).

Solomou, G. D., & Koutsomitropoulos, D. A. (2014). *Support of SKOS Vocabularies in the DSpace Digital Repository System*. (May).

Solomou, G., & Papatheodorou, T. (2010). The use of SKOS vocabularies in digital repositories: The DSpace case. *Proceedings - 2010 IEEE 4th International Conference on Semantic Computing, ICSC 2010*, (October 2010), 542–547. <https://doi.org/10.1109/ICSC.2010.83>

Svenonius, E. (2003) *Design of controlled vocabularies*. Los Angeles; University of California

Tongco, M. D. (2007). Purposive Sampling as a Tool for Informant Selection. *A Journal of Plants, People, and Applied Research*.

## Appendices

### Appendix 1: Structured interview guides

#### 1. Interview Schedule for Library Staff

##### I. Opening

- A. My name is ..... and I am a fourth year student in the School of Education under the Department of Library and Information Science working on a project where are investigating the feasibility of integrating controlled subject vocabulary sets in the UNZA Institutional Repository.
- B. The purpose of this interview is to establish the extent to which controlled subject vocabulary sets are used in tagging digital objects in the UNZA IR and how the integration of these controlled subject vocabulary sets into DSpace is done. It is hoped that this information will help create awareness about the UNZA IR as to its importance and how best we could improve the visibility of the university's intellectual output.
- C. The interview should take no more than 10 minutes. Are you available to answer these questions at this time?

**Transition:** Let me start by asking you some questions about what exactly you do here.

##### II. Body

- A. Do you mind telling me your name and the position you hold in the library?
- B. How do you tag elements in the IR?
- C. How long it takes to create metadata for a typical digital object
- D. What controlled vocabulary sets do you use?
- E. How would you describe the submission workflow into the IR?
- F. Have you faced any challenges in handling the IR?

Yes                       No

- G. If, yes to Q5, have you done anything to try to overcome these challenges?

**Transition:** It has been a pleasure interviewing you. Let me just quickly summarize what I have recorded during the interview.

##### Closing

- A. Summary
- B. I highly appreciate the time you took for this interview. Is there any other information you would like to share that may prove useful to this project?
- C. I should have all the information I need for now. However, would it be okay to call or visit you at a later time in case I have more questions? Thank you once again.

## **2. Interview Schedule for Faculty Staff**

### **I. Opening**

- A. My name is ..... And these are my colleagues [Mention each of them by first and last name]. We are fourth year students working on a project related to the Institutional Repository; therefore, we thought it would be appropriate to have an interview with you to get better informed on the subject. However, before we proceed, could you sign this consent form to show that you are willing to have this interview? [Present Consent form]
- B. We hope to use this information to bring to the attention of the University's populace and those outside, the importance of the integrating subject controlled vocabulary sets in the University of Zambia Institutional Repository.
- C. This interview should take no more than 20 minutes. Shall we begin?

### **II. Body**

- A. What position do you hold within the faculty? Could you briefly outline what you do, precisely?
- B. How long have you worked at the University?
- C. We understand that you do write scholarly papers and articles. Where do you mostly publish literature related to your field of study? Where is the most relevant literature in your field of study located?
- D. Do you have standard phrases and/or words specific to your particular field that you associate with the papers you publish? For example, in medicine, they use Medical Subject Headings such as Disease to associate to papers talking about various diseases.
- E. Would you welcome the idea of self-archiving your work into the University's digital repository?

- F. If your answer to the previous question is no, could you give a reason why?
- G. If yes, do you think it would increase the visibility of your work to members of the institution and to those outside?

**Transition:** It has been a pleasure interviewing you. Let us quickly run through a summary of this interview.

### III. **Closing**

- A. [Give summary]
- B. We highly appreciate the time you took for this interview. Is there any other information you would like to share that may prove useful to this project?
- C. We should have all the information we need for now. However, would it be okay to call or visit you at a later time in case we have more questions? Thank you for taking the time to participate in this interview.