Automatic Classification of Digital Objects for Improved Metadata Quality of Electronic Theses and Dissertations in Institutional Repositories

Lighton Phiri

Department of Library and Information Science,
University of Zambia,
Lusaka, Zambia
E-mail: lighton.phiri@unza.zm

Abstract: Higher Education Institutions typically employ Institutional Repositories (IRs) in order to curate and make available Electronic Theses and Dissertations (ETDs). While most of these IRs are implemented with self-archiving functionalities, self-archiving practices are still a challenge. This arguably leads to inconsistencies in the tagging of digital objects with descriptive metadata, potentially compromising searching and browsing of scholarly research output in IRs. This paper proposes an approach to automatically classify ETDs in IRs, using supervised machine learning techniques, by extracting features from the minimum possible input expected from document authors: the ETD manuscript. The experiment results demonstrate the feasibility of automatically classifying IR ETDs and, additionally, ensuring that repository digital objects are appropriately structured. Automatic classification of repository objects has the obvious benefit of improving the searching and browsing of content in IRs and further presents opportunities for the implementation of third-party tools and extensions that could potentially result in effective self-archiving strategies.

Keywords: Digital Libraries; Dublin Core; OAI-PMH; document classification; automatic classification; digital objects; metadata quality; Electronic Theses and Dissertations; ETDs; Institutional Repositories; self-archiving.

Reference to this paper should be made as follows: 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*, Vol. 14, No. 3, pp.234–248. DOI: 10.1504/IJMSO.2020.112804

Biographical notes: Lighton Phiri is a Lecturer and Researcher in the Department of Library and Information Science at The University of Zambia. He was awarded a PhD in Computer Science at the University of Cape Town. His broad research interests lie within the areas of Digital Libraries and Data Mining. He also has on-going research interest in Information and Communication Technologies for Development (ICT4D) and Technology-Enhanced Learning.

1 Introduction

Institutional Repositories (IRs) (Lynch, 2003) are a specialised type of Digital Libraries (DLs) that are fundamentally used to store, manage and facilitate access to digital objects (Arms, 1995, Arms et al., 1997). Higher Education Institutions (HEIs) typically use IRs to make available scholarly research output that they produce. While there is a broad range of scholarly research output that is deposited into IRs, they generally include pre-prints and post-prints of peer-reviewed journal articles, conference proceedings, books, book chapters, technical reports and Electronic Theses and Dissertations (ETDs) produced by graduate students. Effectively, storing and making available scholarly research output using IRs increases the online visibility of HEIs, a key criterion used when ranking HEIs (Ioannidis et al., 2007). More importantly, online visibility ensures that crucial results of important studies are easily accessible by relevant stakeholders, such as researchers, the Government and Non-profit Organisations.

Ingestion of digital objects in IRs is done either through self-archiving (Harnad, 2001), with the manuscript authors tasked with the responsibility of depositing the manuscript or by a central authority, typically the Library. In both instances, there is the potential to misclassify digital objects by way of depositing them in the wrong collection and, more importantly, tagging them with non-subject specific subject descriptions. In the case of The University of Zambia (UNZA), this has been worsened due to the fact that self-archiving of ETDs is non existent and the Library only has two individuals responsible for the

ingestion of IR objects. Analyses conducted in prior work (Phiri, 2018) highlighted the lengthy turnaround time between submission of scholarly research output and eventual ingestion into the IR. In addition, the lack of use of subject-specific controlled vocabulary sets was noted as being a major concern since it compromises the discoverability of digital objects. Furthermore, we outlined the adverse effect this has on downstream service providers, such as the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog[1] and the Open Access Theses and Dissertations portal[2], all of which automatically harvest metadata from IRs that also function as data providers.

This paper is aimed at demonstrating the feasibility of automatically classifying IR digital objects using the minimum possible input expected from manuscript authors—the ETD manuscript. Incidentally, there is additional information that could potentially be supplied together with the ETD manuscript, for instance meta information of the unit associated with the author and additional contributors, e.g. Advisors and/or Supervisors. The motivation for working towards the automatic classification of ETD digital objects is three-fold:

1. To devise a mechanism for reclassifying digital objects already ingested into the repository.

2. To set the stage for the implementation of tools that could potentially leverage automatic classification of digital objects in order to guarantee accurate classification of digital objects, using standard tags.

3. To reduce the time taken to ingest digital objects into IRs.

The main contributions of this work are as follows:

- Identification of a core set of features, extracted from the minimum possible input—ETD manuscript—provided by ETD authors.

- Classification models for automatically classifying ETD types, associated IR collections and, subject headings for ETDs.

- Demonstration of how standardised controlled vocabulary sets can be associated with ETDs.

This paper is organised as follows: Section 2 presents background information and existing literature related to this work. Section 3 outlines the methodology, in which emphasis is placed on describing core features identified for implementing the classification models. The details of experiments conducted and the corresponding discussion are outlined in Section 4 and Section 5, respectively. Finally, Section 6 presents the conclusion.

2 Related Work

2.1 Institutional Repositories

IRs are domain-specific DL platforms that are designed and implemented to store scholarly research output. Fundamentally, the design of the repository component of DL platforms is done in a manner that facilitates efficient and effective storage and retrieval of two key aspects of digital objects: metadata and bitstreams (Arms et al., 1997). The metadata is typically stored in in a relational database, due to its consistent structure, while the bitstreams are ideally stored on the filesystem. The associative relationship of metadata and bitstreams is typically performed during ingestion of digital objects, during which descriptive, structural and administrative metadata is specified through a submission workflow. The descriptive and administrative metadata generally makes use of standard metadata schemes such as Dublin Core (Weibel et al., 1998), while the structural metadata is application specific and typically determines how the digital objects are organised and presented in the IR. For most IR platforms, organisation takes the form of a hierarchical structure comprising of collections associated with faculties, deparments and units of the institution.

However, the IR submission workflow is generally an error-prone and time-consuming exercise, primarily due to the large amount of structural and descriptive metadata elements that need to be associated with the digital object. For instance one of the most popular IR open source platforms, DSpace, has a submission process that comprises of a series of six steps in its current stable release (DSpace, 2018*b*).

2.2 Digital Object Metadata

As outlined in Section 2.1, digital objects in IRs are represented by the bitstreams—the content associated with the digital objects—and metadata—the metadata information that provides additional contextual information about the digital object.

2.2.1 Types of Metadata

Riley (Riley, 2017) categorises metadata that can be associated with digital content into three broad groups: administrative metadata, descriptive metadata and structural metadata, based on their relative role.

Administrative Metadata. Administrative metadata is aimed at providing information used for managing and processing digital objects, in order to facilitate the long-term preservation of digital objects. The details of this information include technical details about the creation of the digital objects, access control details for specifying access rights associated with the digital objects and details related to rights management. While some of the metadata elements are explicitly provided by the

users during the creation of the digital object, other elements are automatically created by the application. For instance, most IRs will automatically create the date when the digital object was ingested into the repository and the system user responsible for the ingestion of the digital object.

Descriptive Metadata. The primary role of descriptive metadata is the enable effective discovery of digital objects, typically through search and browse services integrated with DLs. The descriptive metadata is generally encoded using a pre-defined metadata scheme and stored in a database management system. While textual IR digital object bitstreams that are born digital can easily be indexed using full-text features available in most IR platforms, some digital objects that are stored in different formats rely on descriptive metadata for their effective discovery.

Structural Metadata. Structural metadata is primarily used to define relationships between digital objects and also to facilitate effective presentation of digital objects, when rendered. Most IRs will present digital objects using container structures for organising related objects, and structural metadata plays a crucial role of associating digital objects to container structures. For IRs used in HEIs, the container structures will generally be associated to faculties, departments and other units that are associated with them.

This work focuses on the metadata used to facilitate discovery of digital objects—descriptive metadata—and the structural metadata used to organise digital objects in IRs.

2.2.2 Metadata Quality in Repositories

The quality of metadata associated with digital objects has been a topic of discussion due to its role is facilitating effective preservation and discoverability of of digital objects. Park states that the quality of metadata is a reflection of the metadata's ability to perform core bibliographic functions of discovery, use, provenance, currency, authentication, and administration (Park, 2009). Park further highlights that common metrics used to assess metadata quality in literature include accuracy, completeness and consistency. These core function can all be mapped on to Riley's broad metadata classifications (Riley, 2017).

Numerous prior work has extensively investigated how the metadata quality in repositories can be improved. Currier et al. used findings from an analysis of learning object repositories to identify errors by untrained resource creators and the lack of use of authority control and subjects as being the major issues associated with quality of metadata (Currier et al., 2004). A study of the Dryad research data repository by Rousidis et al. identified major problems with Creator, Data and Type metadata elements (Rousidis et al., 2014*b*,*a*). In a follow-up study, Balatsoukas et al. (Balatsoukas et al., 2018) performed a descriptive analysis of subject metadata in the Dryad research data repository using SQL queries and uncovered additional quality problems largely attributed to the lack of use of controlled vocabularies.

Some potentially viable techniques proposed to address quality issues include the use of authority control, metadata augmentation and automatic generation of metadata elements. Hillmann hails terminology services and adherence to standards as being crucial aspects for facilitating improved metadata quality (Hillmann, 2008). The terminology services comprise of services for applying vocabularies, while standards ensure consistency of metadata elements. In a survey of DL developments, Tani et al. categorise approaches to addressing metadata quality issues into the following four groups (Tani et al., 2013):

- Metadata guidelines, standards and application profiles—The use of agreed policies and guidelines for characterising resources.
- Metadata evaluation approaches—Computer assisted evaluation strategies for assessing quality dimensions.
- Semi-automatic metadata generation approaches—Combining human approaches with software facilities to promote the creation of metadata.
- Metadata cleaning, enhancement and augmentation approaches—Repairing existing metadata elements.

In this work, we build on prior work (Phiri, 2018) that is aligned with metadata evaluation approaches to devise semi-automatic metadata generation techniques using supervised machine learning.

2.2.3 Summary

The metadata elements associated to digital objects are generally manually prepared by qualified staff and subsequently associated with digital object bitstreams during ingestion of digital objects into IRs. IR platforms usually implement multi-step workflows that facilitate the ingestion process. However, the process of deriving metadata and associating it to bitstreams is time consuming and error prone.

This work proposes to reduce the time spent preparing metadata, thus making the ingestion process more efficient and, to reduce the errors resulting from manual preparation of metadata. This will be accomplished through the automatic classification of digital object structural metadata elements used to define container structures linked to digital objects and, additionally, descriptive metadata elements used to facilitate the discoverability of digital objects.

2.3 Controlled Vocabulary Sets

A key feature of IRs is facilitating effective discoverability of content through searching and browsing functionalities (Arms et al., 1997). While it is typically the case that full-text searching of bitstreams is possible, the searching and browsing is done on metadata elements associated with digital objects. In order for IRs to cluster related digital objects together, controlled vocabulary sets are used. The vocabulary sets provide a mechanism for presenting a restricted set of terms during ingestion of digital objects into IRs.

The use of subject-specific controlled vocabularies in large academic databases provide use cases for why their use is important. For instance, the arXiv repository[3] uses subject classifications for content ingested into the repository. Digital objects indexed in the ACM Digital Library[4] are normally tagged using the ACM Computer Classification System (CCS) (Coulter, 1997). Some PubMed Central[5] articles are tagged with the Medical Subject Heading (MeSH) classes (Dhammi & Kumar, 2014).

Existing literature highlights effective discoverability (Phillips et al., 2019) and improved interoperability (Gries et al., 2018, Schirrwagen et al., 2016) as being the key advantages of using controlled vocabularies.

UNZA currently makes use of the Library of Congress Subject Headings (LCSH) vocabulary. However, ETDs that are stored in IR are authored by postgraduate students from different faculty disciplines. Specifically, UNZA comprises of 13 faculties, each with different departments that are associated with discipline specific controlled vocabularies. For instance, the School of Natural Sciences is composed of the following departments: Biological Sciences, Chemistry, Computer Science, Mathematics, Physics and Geography and Environmental Studies. Each of these departments can potentially be associated with a wide variety of controlled vocabularies.

2.4 Data Standards and Interoperability Protocols

Interoperability is a computer system's ability to be interfaced with other external system services through the standardised use of pre-defined data formats and communication protocols. Suleman (Suleman, 2011) states that in the context of DLs, interoperability promotes openness, a key philosophy mandated by the Open Access movement.

DLs generally encode bitstreams using data formats, such as JPEG[6], PNG[7] and PDF/A[8], which facilitate the long term preservation of data. Metadata is also encoded using well-established international standards such as Dublin Core. The use of international standards ensures that other tools and services are easily able to use and integrate the data.

DLs employ communication protocols to provide auxiliary services for facilitating the core functionalities associated with repositories—ingestion, management, search and browsing of digital objects. For instance, communication protocols such as SWORD (Lewis, 2012) are used for remote ingestion of content, while Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al., 2002) and Open Archives Initiative Object Reuse and Exchange (OAI-ORE) (Lagoze et al., 2008) are used for harvesting digital object metadata and bitstreams, respectively.

In this work, the OAI-PMH and OAI-ORE protocols were used to harvest data used for creating datasets for implementing classification models. The OAI-PMH protocol provides the following six verbs—GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords and ListSets—that are used to interact with repositories (Lagoze et al., 2002).

2.5 Self-Archiving Using Institutional Repositories

Self-archiving is defined as the process that enables a manuscript author to deposit electronic publications in order to facilitate open access (Harnad, 2001). While there has been tremendous support for self-archiving, especially when it comes to IRs, there still remain challenges. In a study conducted by Katayoon and Abrizah in eight universities in Malaysia, it was revealed that ingestion of digital objects in to repositories is mainly done by Librarians rather than authors (Katayoon & Abrizah, 2010).

Existing literature has identified several factors that affect self-archiving. In attempting to uncover the motivations associated with self-archiving behaviour and factors that makes faculty reluctant to self-archive, Kim proposes a model of factors that influences self-archiving behaviour. Kim cites "Additional time and effort" and "Technical skills" as having negative and positive associations, respectively (Kim, 2010, 2011). Other researchers have highlighted experience as being correlated with self-archiving. Xia supports the experience argument by noting that it has a positive correlation to self-archiving (Xia, 2007). In a study of awareness and self-archiving practices of Kenyan academics in five universities, the findings suggest that awareness and attitude of academics is low (Chilimo, 2016), with the vast majority supporting mandatory open access policies. The low uptake in self-archiving is supported by other existing studies (Baro et al., 2018), necessitating the need for techniques that emphasise the use of automatic classification of IR objects.

While self-archiving is hailed as a solution to ensuring timely ingestion of content, the automatic classification of IR digital objects presents opportunities for implementing tools that could potentially reduce errors and the time spent ingesting content. More importantly, automatically classifying digital objects has the potential to complement self-archiving initiatives.

UNZA still has to grapple with this issue, as most content ingested into the IR are deposited by Library staff. Incidentally, while, there are some faculty staff at

UNZA that self-archive pre-prints of their publications, a significant proportion are unable to and, all ETDs are exclusively deposited into the repository by Library staff. The automatic classification approach proposed in this paper is, in part, aimed at reducing the amount of time spent depositing content into IRs. Furthermore, the automatic classification of digital objects presents opportunities to implement software tools that are usable and easy to use.

2.6 Supervised Machine Learning

The continued increase of the rate at which data is being generated has seen a rise in the application of machine learning techniques. These machine learning techniques are broadly categorised into three main types—supervised machine learning, unsupervised machine learning and reinforcement learning—and are fundamentally used for making predictions and to identify patterns in data (Abu-Mostafa et al., 2012). Of the three machine learning techniques, supervised machine learning is the most widely used and generally involves designing algorithms that learn by example.

Learning from examples in supervised machine learning techniques relies on labelled datasets that provide a mapping between input variables and expected output variables. The learning algorithm uses the mapping between labelled input and output variables to generate a learning function that can be used to predict output variables for new input variables. In addition, supervised machine learning techniques can be grouped into two main categories: regression and classification. Regression problems focus on predicting nominal output variables, while classification problems are aimed at predicting categorical output variables.

This work, in part, uses supervised machine learning to implement classification models for predicting digital object IR structural and descriptive metadata elements.

2.7 Automatic Classification of Documents in Digital Libraries

The area of document classification has been extensively explored, although much focus has been restricted to topic modeling techniques (Blei et al., 2003). A common technique used during text classification involves transforming documents in a corpus into a document term matrix that indicates the relative importance of terms in individual documents, relative to the entire corpus (Sebastiani, 2002).

Prior work on document classification in IRs has primarily focused on explicitly classifying text documents with the goal of enhancing descriptive metadata for improved discoverability of scholarly research output.

Al-Digeil et al. propose an automated technique for generating metadata for IR digital objects using machine learning (Al-Digeil et al., 2007). Support Vector Machines were found to be effective in identifying descriptive metadata elements. While this study shares similarities with our proposed approach, by way of using text features to classify IR objects, their focus is on the automatic generation of metadata restricted to a broad list descriptive metadata: title, author, affiliation, address and abstract. Our work aims to compound descriptive metadata with structural metadata. Furthermore, our focus is on a subset of IR digital objects: ETDs.

Caragea et al. propose the use of structural features aimed at classifying different document types (Caragea et al., 2016). Their experiments suggest that the structural features outperform bag-of-words and URL features. While this approach is similar to the approach proposed in this paper, our focus is on the classification of ETD types, subject categories and collection structures associated with the ETDs. Furthermore, while the approach proposed in this paper uses similar features—e.g. number of pages in document—the use of such features is as a direct result of guidelines associated with the case study context: UNZA.

Charalampous and Knoth propose an approach to document classification aimed at addressing inconsistencies of metadata in repositories in order to provide seamless search and recommender services (Charalampous & Knoth, 2017). They propose the use of text features in order to classify research papers, slides and ETDs. Numerous additional literature focused on document classification approaches that emphasise the use TF-IDF representations of text features as input variables (Kowsari et al., 2019, Hafnan & Mohan, 2018).

While the study uses text features, the focus of this work is different in that it aims to not only classify descriptive metadata—the ETD type and subject categories associated with ETDs—but also structural metadata for identifying container structures that the ETDs are associated with.

3 Methodology

As earlier outlined, the goal of this study was to implement classification models for automatically classifying ETDs. Specifically, three classification models, outlined in Table 1, were implemented in order to automatically associate structural metadata and descriptive metadata to ETDs.

Table 1 Metadata Classification Models

Aspect	Metadata	Classification Model
ETD Type	Descriptive	Binary
ETD Subject	Descriptive	Multi-class
ETD Collection	Structural	Multi-class

The study was conducted by following the standard CRISP-DM data mining model (Wirth & Hipp, 1995), with all of the six phases utilised as follows:

- Business Understanding—Prior work conducted was, in part, aimed at understanding the role of relevant stakeholders within the ETD submission and ingestion workflows. Section 3.1 briefly outlines the roles of the various stakeholders.
- Data Understanding—ETD metadata and bitstreams were analysed to gain in-depth understanding of the elements associated with ETD digital objects.
- Data Preparation—The data preparation process for text features employed all common text preprocessing techniques: removal of stopwords, punctuation marks and numbers, stemming, and handling of null values.
- Modeling—The key features outlined in Section 3.2 were identified and incorporated into the model implementation phase.
- Evaluation—The classification models were evaluated to assess their relative effectiveness by measuring the accuracy of standard Machine Learning estimators. In addition, the effectiveness of feature combinations was assessed, as outlined in Section 4.
- Deployment—Application Programming Interface (API) endpoints were implemented for the classification models to facilitate the integration of the models with third-party tools and services.

3.1 UNZA ETD Submission Workflow

As earlier mentioned, the approach proposed uses the minimum input expected from ETD manuscript authors—the ETD manuscript. The Directorate of Research and Graduate Studies (DRGS) at UNZA has set up guidelines (Directorate of Research and Graduate Studies, 2015) of the end-to-end processes associated with manuscript preparation, examination and submission as shown in Figure 1 and outlined below.

Step #1: Manuscript Preparation. Once the student successfully presents their work during an oral examination session, the final version of the manuscript is prepared, incorporating suggested changes and corrections from examiners. The manuscript conforming to set guidelines is submitted to DRGS as a single electronic copy on a compact disk, in addition to two printed copies.

Step #2: ETD Verification and Validation. DRGS verifies and validates the ETD manuscripts to ensure that they conform to prescribed guidelines (Directorate of Research and Graduate Studies, 2015). Once the verification and validation is completed, the manuscripts are sent to the Library for archiving and ingestion into the IR.

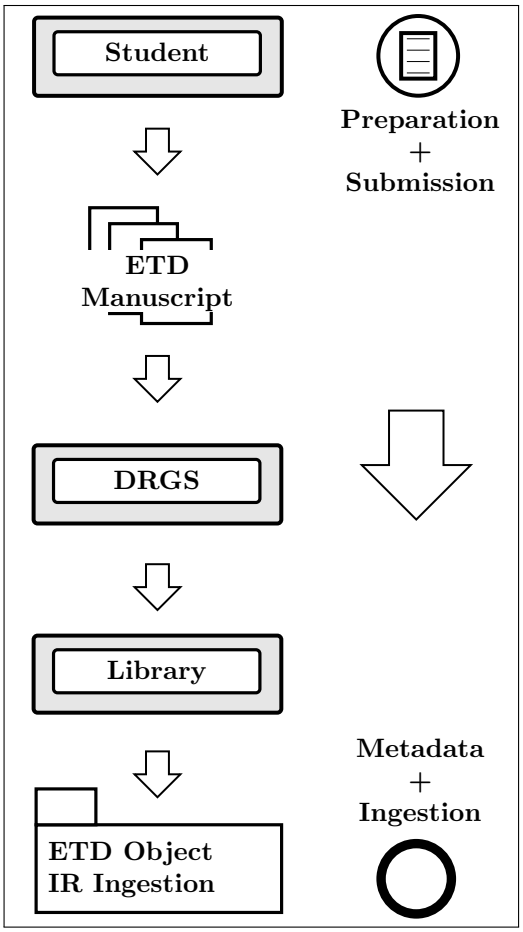


Figure 1: UNZA ETD Submission Workflow

Step #3: ETD Ingestion. The Library prepares Dublin Core encoded descriptive metadata associated with each ETD manuscript by reading through the titles and abstracts associated with each ETD. The ETD is subsequently deposited into the IR by specifying the appropriate IR community and collection associated with the manuscript.

Listing 1 shows the hierarchical structure of UNZA's IR. The communities and collections are a mirror of faculties and departments at UNZA.

Listing 1: UNZA's Hierarchical IR Structure

```
Agricultural Science [Community]
    + Crop Science [Collection]
    + [...]
    + Animal Science [Collection]
        + Digital Object [Article]
            + Metadata [Dublin Core]
            + Bitstream [PDF]
[...]
Theses and Dissertations [Community]
    + Education [Collection]
    + [...]
    + Engineering [Collection]
        + Digital Object [ETD]
            + Metadata [Dublin Core]
            + Bitstream [PDF]
```

3.2 Feature Engineering

The automatic classification approach proposed is aimed at classifying ETDs ingested into HEI IRs. While most features proposed in this work are generic, some of are specific to UNZA, as they are based on UNZA's postgraduate regulations, as outlined in Section 3.1.

3.2.1 Feature Extraction Using Author Supplied Information

Using the information supplied from the user, implicitly—using descriptive metadata encoded with bitstreams during the ETD submission workflow—and explicitly—using the PDF manuscript—, we propose to extract features that could potentially facilitate the classification of the collection hierarchical structure where the ETD should be deposited and the type of ETD. Furthermore, we use some of the features to predict subject classes using a trained model from a well-established repository, as described in Section 3.2.2.

Digital Object Bitstreams. UNZA guidelines for preparing thesis and dissertation manuscripts specifies the the textual content that must be present on the title page and, additionally, there are limits placed on the maximum number of pages for a Master's dissertation and Doctoral thesis (Directorate of Research and Graduate Studies, 2015).

- Master's Manuscripts—The maximum number of pages allowed for Master's theses and dissertations is 60,000—approximately 190 pages, using one-and-a-half spacing.
- Doctoral Manuscripts—The maximum number of pages allowed for Doctoral theses is 100,000—approximately 320 pages, using one-and-a-half spacing.

Listing 2: Sample ETD Manuscript Title Page

```
1   A CROSSECTIONAL STUDY OF FACTORS
2  CONTRIBUTING TO MODERATE TO SEVERE
3  POST OPERATIVE PAIN AFTER A LAPAROTOMY
4
5
6                  By
7           USHMABEN PATEL
8        (BSc.HB, MB.ChB)
9
10
11  A dissertation submitted to University X
12 in partial fulfilment of the requirements of
13    the degree of Master of Medicine in
14       Anaesthesia and Critical Care
15
16
17
18              University X
19                  City
20                  2017
```

For both Master's and Doctoral manuscripts, the title page is expected to explicitly state the title of the degree and the field of the degree, as shown in Listing 2—Lines 11–14.

This paper proposes to use the text on the title page as text features that would help determine the type of ETD. Furthermore, the merged PDF manuscript could easily be used to determine the total number of pages for manuscripts, as shows in Listing 3—Line 11.

Listing 3: Sample ETD PDF Document Metadata

```
1  Author:          User
2  Creator:         Microsoft Word 2010
3  Producer:        Microsoft Word 2010
4  CreationDate:    Wed Nov  8 21:18:34 2017 CAT
5  ModDate:         Wed Nov  8 21:18:34 2017 CAT
6  Tagged:          yes
7  UserProperties:  no
8  Suspects:        no
9  Form:            none
10 JavaScript:      no
11 Pages:           61
12 Encrypted:       no
13 Page size:       595.32 x 841.92 pts (A4)
14 Page rot:        0
15 File size:       990008 bytes
16 Optimized:       no
17 PDF version:     1.5
```

Digital Object Descriptive Metadata. The ingestion of digital objects into IRs typically involves a workflow process that facilitates the tagging of digital objects with descriptive metadata. UNZA's IR runs a DSpace instance, which is loosely based on the qualified Dublin Core metadata scheme (Weibel et al., 1998). It is worth mentioning that the base metadata scheme configured with DSpace can be cross-walked to a different scheme: for instance, Dublin Core can be cross-walked to ETD-ms (The Networked Digital Library of Theses and Dissertations, 2015). Listing 4 shows a sample ETD metadata record, encoded using Dublin Core.

In addition to the descriptive metadata elements, DSpace holds two additional types of metadata: administrative and structural metadata. The administrative metadata includes preservation, provenance and authorisation metadata elements, while the structural metadata that specifies the nested structures where the digital object is ingested, once deposited into the IR (DSpace, 2018*a*).

Out of all the Dublin Core elements and additional structural elements, only the Title and Description elements comprise of information supplied by manuscript authors. The Title element (Lines 14–19 in Listing 4) is used to encode the manuscript title, while the Description element (Lines 39–53, in Listing 4) is used to encode the manuscript abstract. When used in isolation and/or combined together, the Title and Abstract provide useful data that can be used as input features. The text features can easily be used to derive TF and TF-IDF features.

3.2.2 Learning from Other Repositories

While it is fairly common to employ topic modeling techniques such as Latent Dirichlet Allocation (Blei et al., 2003) to automatically generate tags for textual documents, our proposed approach uses a trivial, yet effective, technique of using well-established DLs in order to derive appropriate subject tags for ETDs.

In order to demonstrate the proposed subject classification approach, the MeSH vocabulary is used as the base knowledge organisation system for identifying appropriate subjects. The National Library of Medicine (NLM) MeSH thesaurus is used to train a model that uses text features to predict MeSH Class 1 Descriptors (National Library of Medicine, 2019*b*), and then use the trained model to predict appropriate subject classes for ETDs from UNZA's Faculty of Medicine. Specifically, annual MEDLINE/PubMed citation records (National Library of Medicine, 2019*a*) are used to train a classification model, as outlined in Section 4.1.3.

ETDs from UNZA's Faculty of Medicine were used because a large proportion of ETDs in UNZA's IR are from that faculty. In addition, most medical fields use the MeSH classification system to tag their publications.

3.3 Classifier Training Labels

In order to build classification models for automatically determining the ETD collection, subject and type, training labels were prepared as follows:

- ETD Collection Labels—Structural metadata corresponding to the ETD communities and collections was used to identify collection labels. In Listing 4, Lines 7–8 correspond to collections associated with the sample ETD, while Lines 9–10 correspond to the communities associated with the ETD.

- ETD Subject Labels—Parent MeSH subjects associated digital objects from MEDLINE/PubMed citation records were used as labels to train a classifier that was used to predict subject tags for ETDs originating from the Faculty of Medicine at UNZA.

- ETD Type Labels—Textual context on titles pages was used to determine the type—Master's or Doctoral—of manuscript. Specifically, the text that was used is represented by Lines 11–14 in Listing 2.

3.4 Summary

In this section, an approach to extract features from IR digital objects has been outlined. Specifically, the features are extracted from digital object bitstreams—the PDF file—and corresponding metadata—Dublin Core encoded records.

Listing 4: Sample ETD Structural and Descriptive Metadata

```
<record>
 <header>
  <identifier>
   oai:dspace.unza.zm:123456789/5701
  </identifier>
  <datestamp>2019-08-19T12:31:59Z</datestamp>
  <setSpec>com_123456789_5087</setSpec>
  <setSpec>com_123456789_18</setSpec>
  <setSpec>col_123456789_6021</setSpec>
  <setSpec>col_123456789_83</setSpec>
 </header>
 <metadata>
  <oai_dc:dc>
   <dc:title>
   A crosssectional study of factors
   contributing to moderate to severe
   post operative pain after a
   laparotomy
   </dc:title>
   <dc:creator>Patel, Ushmaben</dc:creator>
   <dc:subject>
   Laparotomy--Zambia
   </dc:subject>
   <dc:subject>
   Obstetric surgical procedures--Laparotomy
   --Zambia
   </dc:subject>
   <dc:subject>
   Urologic surgical procedures--Laparotomy
   --Zambia
   </dc:subject>
   <dc:subject>
   Gynecologic surgical procedures--Laparotomy
   --Zambia
   </dc:subject>
   <dc:description>
   Thesis
   </dc:description>
   <dc:description>
   Having pain relief is a basic human right.
   This thesis provides a descriptive profile
   of pain before and during the first 72 hrs
   after a laparotomy at University Teaching
   Hospital, Lusaka, Zambia, from July 2014 to
   January, 2015. The objective was to identify
   independent risk factors associated with
   moderate to severe pain after laparotomy.
   [...]
   A midline incision was more painful than
   transverse at 24 hrs, 48  hrs and 72 hrs
   post-operative period.
   [...]
   </dc:description>
   <dc:date>2019-01-30T08:03:29Z</dc:date>
   <dc:date>2019-01-30T08:03:29Z</dc:date>
   <dc:date>2017</dc:date>
   <dc:type>Thesis</dc:type>
   <dc:identifier>
   http://dspace.unza.zm/handle/123456789/5701
   </dc:identifier>
   <dc:language>en</dc:language>
   <dc:format>application/pdf</dc:format>
   <dc:publisher>
   The University of Zambia
   </dc:publisher>
  </oai_dc:dc>
 </metadata>
</record>
```

In addition, the section outlined how external repositories can be leveraged to train classification

models for classifying document subject categories based on standardised controlled vocabulary sets. Table 2 shows a summary of features extracted.

Table 2 Feature Extraction from ETD Manuscript

Feature	Description
TitlePage	Bag-of-words: title page text
NumberPages	ETD length in pages
ETDTitle	Bag-of-words: ETD title
ETDAbstract	Bag-of-words: ETD abstract

4 Evaluation

All the experiments were conducted on a standalone LENOVO® IdeaPad 320, with an Intel® Core™ i7-8550U (CPU @ 1.80GHz), using 12 GB RAM, and running Ubuntu 18.04.3 LTS[9].

4.1 Datasets

Datasets used for conducting experiments were prepared using data harvested from UNZA's IR and, additionally, annual MEDLINE/PubMed citation records (U.S. National Library of Medicine, 2018).

4.1.1 OAI-ORE Harvested PDF Documents

ETD PDF manuscripts were harvested from UNZA's IR using the OAI-ORE specification (Lagoze et al., 2008). The distribution of ETDs in the various faculties at UNZA are shown in Table 3.

Two separate datasets were prepared as follows, using the harvested PDF manuscripts:

Table 3 ETD Distribution in UNZA IR

Faculty	Master's	Doctoral	Unclassified	Total
Medicine	636	9	127	772
Education	683	34	35	752
Social Sciences	690	7	46	743
Law	51	2	221	274
Natural Sciences	181	7	43	231
Agricultural Sciences	157	4	34	195
Distance Education	90	–	1	91
Engineering	72	8	6	86
Mines	61	2	12	75
Library	37	1	30	68
Veterinary Medicine	63	4	1	68

Dataset #1—PDF Title Pages. The PDFtk (Steward, 2019) utility was used to create a dataset comprising of title pages, by extracting the first page of the harvested PDF manuscripts.

The title pages were subsequently converted into equivalent text documents, using the pdftotext utility[10]. However, a significant proportion—47.6%—of harvested PDF documents were scanned bitstreams and as such a processing pipeline combining ImageMagic's convert[11] utility and tesseract[12] were used to convert the cover pages to text.

The text documents were then used to build a bag-of-words model consisting of terms on the title page

Dataset #2—Number of Pages. It is common for different portions of manuscripts to be ingested into IRs as separate bitstreams, linked together using the relation Dublin Core element (Weibel et al., 1998). The PDFtk utility was used to merge all digital objects comprising of multiple bitstreams. The pdfinfo[13] utility was subsequently used to determine the total number of pages associated with each manuscript.

4.1.2 OAI-PMH Harvested Metadata

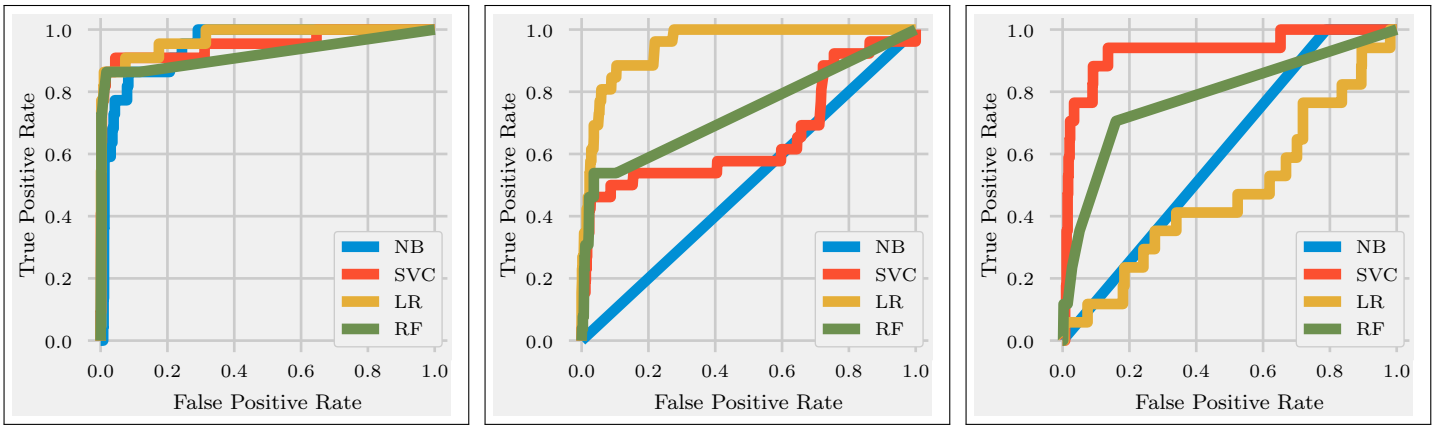
Dublin Core (Weibel et al., 1998) encoded metadata records where harvested from X University's IR using the OAI-PMH standard (Lagoze et al., 2002). While Dublin Core uses a set of 15 elements, the following were utilised:

- Identifier—Used to uniquely identify the ETD manuscripts and link the various datasets.
- Title—Used to extract text features used to build the classification model.
- Description—This is typically used to encode the ETD abstract and was thus used to extract text features used to build the classification model.
- Type—Used for verifying the type of ETD labels for distinguishing Master's dissertations and Doctoral theses.
- SetSpec—Used for labelling the ETDs in order to distinguish the faculty where the ETD was prepared from.

4.1.3 Baseline MEDLINE/PubMed Citation Records

XML encoded MEDLINE/PubMed records were download from the batch of most recent annual baselines (National Library of Medicine, 2019*a*). A dataset was then created to be used for training the model for classifying MeSH headings associated with ETDs.

Table 4 shows a summary of the datasets used for conducting experiments.



(a) Cover Pages (b) Number of Pages (c) Cover & Pages

Figure 2: ROC Curves for ETD Type Binary Classification Model

Table 4 Datasets Used for Experimentation

Dataset	Feature	Records	Size
ETD Title Pages	TF-IDF	3,339	14GB
ETD Total Pages	# Pages	4,086	15GB
ETD Metadata	TF-IDF	3356	13MB
Medline Baselines	TF-IDF	1,263,903	12GB

4.2 ETD Classification Models

Training and testing datasets were created using the hold-out method built within the scikit-learn Python library, with 70% of each dataset used for training and the remaining 30% for testing. Further, the implementation of classification models was also done using the scikit-learn Python library (Pedregosa et al., 2011). The model evaluation process employed various scikit-learn implementations of Logistic Regression, Naive Bayes (Multinomial), Random Forest and Stochastic Gradient Descent.

4.2.1 ETD Type Classification

In order to classify the type of ETD, the number of pages (ETD # Pages) of the ETD and textual content on the title page (ETD Title Page) were used as features. The ETD title page text was transformed into its equivalent TF-IDF representation, while the numeric representation of the number of pages of the ETD was used in its normal form.

4.2.2 ETD Collection Classification

The ETD collection classification model was implemented using two features: the ETD title and ETD abstract. Experiments were conducted to determine the influence of the title (ETD Title), the abstract (ETD Abstract) and combining the title and abstract (ETD Title+Abstract). In all three instances, the TF-IDF representation of the features were used as input.

4.2.3 ETD Subject Classification

In order to demonstrate the feasibility of implementing a subject classification model, ETDs from the Faculty of Medicine were used as a test case. The subject classification model was implemented using annual baseline MEDLINE/PubMed citation records. The citation records titles (Title), abstract (Abstract), MeSH label and combinations of the three were used to determine the most effective features.

5 Results and Discussion

5.1 ETD Type Classification

Figures 2a to 2c and Table 5 show relative performance of the ETD type classification model, using the two different features: ETD Title Page and ETD # Pages. The results suggest that the ETD # Pages feature—using RandomForestClassifier—results in better performance when compared with the ETD Title Page feature. The natural expectation is for the text features on the title to outperform the ETD # Pages, especially that it is possible to have Doctoral manuscripts with pages falling within the Master's manuscript page limit threshold. The reason why the ETD Title Page under-performed is due to the fact that a significant proportion of manuscripts are ingested as scanned bitstreams, making it difficult to extract textual content on the title page. In addition, some manuscripts appear to have been uploaded without title pages.

5.2 Collection Classification

The performance of features used to classify collections are shown in Figure 3 and Table 6. Interestingly enough, using the ETD manuscript abstract results in better performance than the combined effect of the ETD title and abstract (Title+Abstract), with an accuracy score of 82.1%, using SGDClassifier. A possible explanation

Table 5 ETD Type Classification Feature Selection

Feature	Precision	Recall	F1-Score	Accuracy
ETD Title Page	0.99	0.99	0.99	98.9%
ETD # Pages	0.94	0.97	0.95	96.8%
ETD Title & # Pages	0.95	0.97	0.96	97.4%

Table 6 Collection Classification Feature Selection

Feature	Precision	Recall	F1-Score	Accuracy
ETD Title	0.55	0.59	0.57	59.0%
ETD Abstract	0.82	0.82	0.82	82.1%
ETD Title+Abstract	0.82	0.81	0.81	81.2%

for the lower performance of Title+Abstract could be attributed to the common phrases that manuscript authors use in titles.

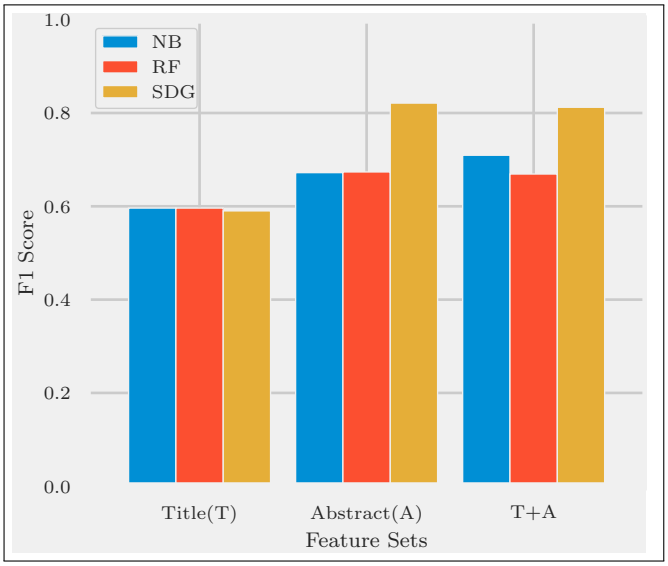


Figure 3: F1 Score: ETD Collection Classification

The results of the classification of the individual multiclasses are shown in confusion matrix in Figure 4 and in Table 7. While the classification results for most fields of study are within acceptable limits, there are misclassifications that are mostly attributed to fields that are similar. For instance, the proportion of Institute for Distance Education (ID) ETDs classified as being part of the Education (ED). A similar explanation can be used for Veterinary Medicine (VM) ETDs classified as being part of the Medicine (MD) collection. A potential workaround could be to use the Title Page text features once the OCR issue with scanned document is resolved. Incidentally, the lower F1-Score values are exhibited by

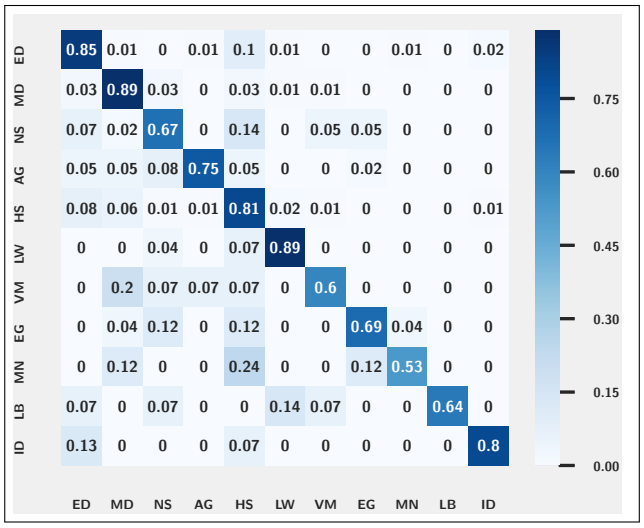


Figure 4: Confusion Matrix: ETD Collection Classification

Table 7 ETD Collection Classification Performance

Collection	Precision	Recall	F1-Score
Education (ED)	0.88	0.87	0.88
Medicine (MD)	0.86	0.92	0.89
Natural Sciences (NS)	0.74	0.48	0.58
Agricultural Sciences (AG)	0.89	0.82	0.86
Social Sciences (HS)	0.80	0.81	0.80
Law (LW)	0.70	0.93	0.80
Veterinary Medicine (VM)	0.67	0.53	0.59
Engineering (EG)	0.77	0.77	0.77
Mines (MN)	0.67	0.59	0.62
Library (LB)	0.90	0.64	0.75
Distance Education (ID)	0.86	0.80	0.83

ETDs from fields with concepts expected to appear in other fields.

5.3 Subject Classification

5.3.1 Model Performance

The performance results of the MeSH subject classification model are shown in Figure 5 and Table 8. When comparing the title, abstract and title+abstract features, the best performing feature was title+abstract, with an accuracy score of 54.6%, using the SGDClassifier. While noticeably low, the accuracy score is comparable to results reported in similar studies involving automatic classification of MeSH subject headings (Mao & Lu, 2017). One possible reason for the poor performance of the Title+Abstract feature could be due to some citation records having missing abstracts. A follow-up study could potentially take a

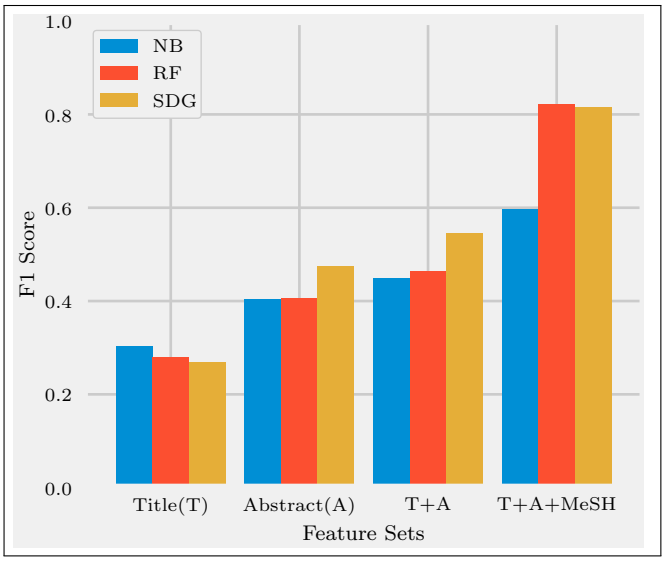


Figure 5: F1 Score: ETD MeSH Classification

Table 8 Subject Classification Feature Selection

Feature	Precision	Recall	F1-Score	Accuracy
Title	0.27	0.20	0.19	30.3%
Abstract	0.42	0.43	0.41	47.5%
Title+Abstract	0.49	0.50	0.48	54.6%
Title+Abstract+MeSH	0.81	0.77	0.77	82.3%

selective strategy of including citation records with complete metadata elements.

In order to enhance the accuracy of the model to predict potential ETD MeSH Class 1 Descriptors, the MEDLINE/PubMed MeSH Descriptors associated with each record were used as complementary input features. The rationale for this decision is that the Descriptors—especially the more specific Class 2, 3 and 4 Descriptors—are more likely to be included in ETD abstracts. The resulting feature (Title+Abstract+MeSH) results in an accuracy score of 82.3%, using the SGDClassifier.

5.3.2 Model Prediction

The trained MeSH subject classification model was used to predict parent MeSH tags associated with ETDs from the Medicine collection and the results are shown in Table 9. It should be noted that only tags associated with at least two ETDs are shown in Table 9.

Using this automated approach to provide subject-specific tags makes it possible for related manuscripts to be associated to each other using browse features implemented within IRs. More importantly, the human effort required to generate subject-specific tags and the likelihood of spelling mistakes are reduced. UNZA has a total of 13 faculties, with each being composed of several departments using a variety of subject-specific controlled vocabularies, making it essential to apply such automated approaches to generate subject-specific labels.

Table 9 MeSH Heading Predictions

MeSH Heading	Parent Tree	No.
Stents	E07.695	350
Health Knowledge	N05.300.150	126
Communication	F01.145	62
Drug Resistance	G07.690.773	39
Algorithms	L01.224	32
Attitude of Health Personnel	F01.100	29
Cognition	F02.463	24
Internet	L01.224.230.110	19
Adaptation	G16.012	11
Cost of Illness	N03.219.151	11
Kidney Transplantation	E04.950.774	7
Clinical Competence	N05.715	6
Tomography	E01.370.350.350	6
Environmental Monitoring	N06.850.460.350	5
Models	N06.850.520.830	5
Decision Making	N04.452	5
Biosensing Techniques	E05.601	4
Dietary Supplements	G07.203	3
Curriculum	I02	3
Drug Delivery Systems	E02.319	2
Aging	G07.345	2
Agriculture	J01	2
Biomarkers	D23	2
Body Mass Index	E05.041.124	2
Postoperative Complications	C23.550	2
Health Status	I01.240	2

5.4 Deployment of Models

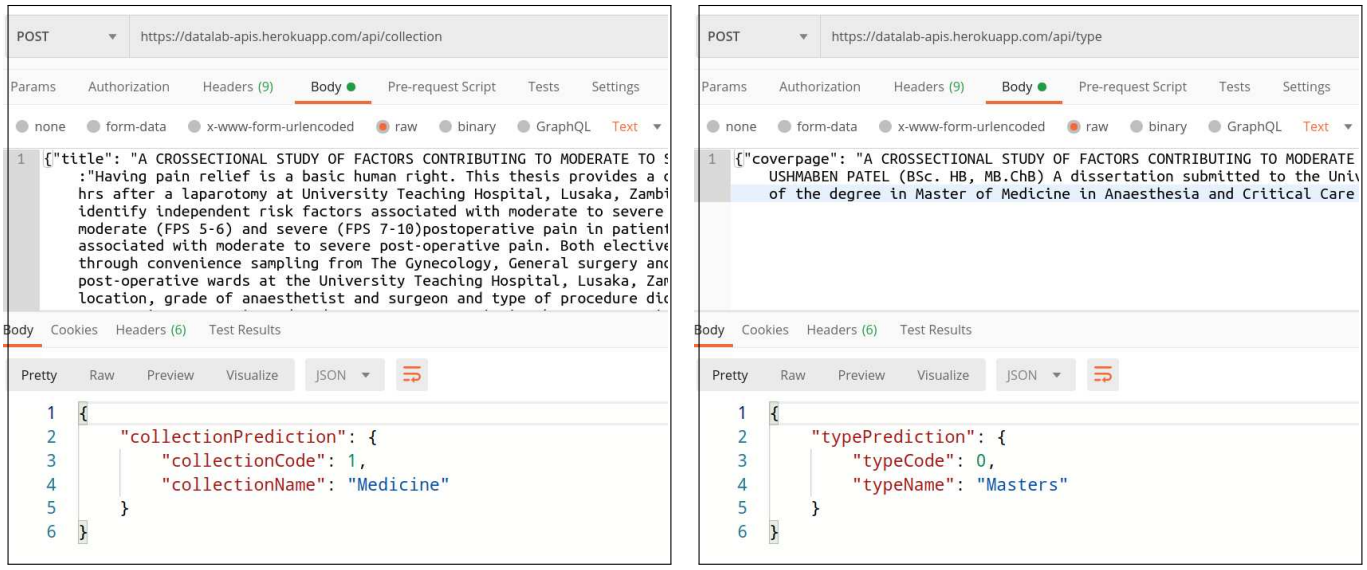
The models are all implemented using offline learning (Ben-David et al., 1997), with their state persisted to disk using the joblib library (Joblib Developers, 2008). Furthermore, to facilitate the implementation of useful tools and services that make use of the models, corresponding APIs have been implemented (Phiri, 2020) using the Python Flask Web framework (The Pallets Project, 2010).

Figures 6a and 6b show sample output from the collection[14] and type[15] models, respectively. The input ETD manuscript for the sample output is shown in Figures 7 and 8.

6 Conclusion and Future Work

6.1 Future Work

This paper was aimed at demonstrating the feasibility of classifying IR digital objects. As part of a broad goal of ensuring that there is increased online visibility of scholarly research output at UNZA, there are a number



(a) Sample IR Collection Classification

(b) Sample Manuscript Type Classification

Figure 6: Sample Output of Flask-based API Endpoints

of open areas, with potential future work taking the form of the following aspects:

- Applying the proposed classification techniques to re-organise the ETDs already ingested into the repository. This would require a comprehensive experimental design involving data that is appropriately cleaned in order to address the poor model performance outlined in Section 5.3.1.

- Exploration of appropriate controlled vocabulary sets to integrate with UNZA's IR submission workflow and subsequently experimenting with multi-label classification of subject categories.

- Identifying additional approaches for comprehensively classifying ETDs using required ETD-ms metadata elements; for instance, automatically generating missing core metadata elements like contributors details

- Implementation of effective software tools and plugins that leverage automatic classification of IR objects, by taking advantage of APIs described in Section 5.4.

- Applying the approach presented in this paper on larger datasets such as the South African National Electronic Theses and Dissertation portal (Webley et al., 2011) and the NDLTD Union Catalog (Suleman, 2012).

- Devising techniques for automatically classifying other IR digital object types such as journal articles and technical reports by building on work already conducted (Caragea et al., 2016).

6.2 Practical Implications

The preparation of ETD metadata, prior to ingestion into IRs and the ingestion process itself are activities that are time consuming and error prone. This is especially the case for HEIs that are under-resourced. Using supervised machine learning techniques outlined in this paper, it becomes possible to reduce errors that result during the preparation of metadata. Furthermore, the time spent ingesting ETDs into IRs can potentially be reduced as previously manual processes are automated.

In essence, in the proposed approach, the role of staff during ingestion would fundamentally involve verifying that the results of automatically classified ETDs are accurate. Using API endpoints described in Section 5.4, the verification and validation process can be incorporated as part of the IR's ingestion workflow. Alternatively, the verification and validation can be implemented as an integrated service of an external service that produces an output that can easily be ingested into the repository.

6.3 Conclusions

This paper outlined a potentially viable way of automatically classifying ETDs in IRs using mandatory information provided by ETD manuscript authors. While classifying the ETD type is achieved to acceptable degrees, classification of collection structures is compromised for closely related fields. The paper demonstrates how external repositories can be used for classifying ETDs using popular controlled vocabulary sets.

With the rapid increase in the number of scholarly publications being generated, it is increasingly becoming important to ensure that digital content is correctly and

comprehensively tagged to facilitate effective searching and browsing of content. The incorporation of machine learning techniques into the ingestion process plays the crucial role of ensuring that the time consuming and error prone tasks are automated, with human users complementing this process through the validation and verification of the end result of the automation process.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. (2012), *Learning from Data: A Short Course*, Kindle edn, AMLBook New York, NY, USA:. [online] `https://work.caltech.edu/textbook.html` (Accessed 25 August 2020).

Al-Digeil, M., Burk, A., Forest, D. & Whitney, J. (2007), 'New Possibilities for Metadata Creation in an Institutional Repository Context', *OCLC Systems and Services* **23**(4), 403–410. [online] `https://doi.org/10.1108/10650750710831547` (Accessed 25 August 2020).

Arms, W. Y. (1995), 'Key Concepts in the Architecture of the Digital Library', *D-Lib Magazine*. [online] `http://www.dlib.org/dlib/July95/07arms.html` (Accessed 25 August 2020).

Arms, W. Y., Blanchi, C. & Overly, E. A. (1997), 'An Architecture for Information in Digital Libraries', *D-Lib Magazine*. [online] `http://www.dlib.org/dlib/february97/cnri/02arms1.html` (Accessed 25 August 2020).

Balatsoukas, P., Rousidis, D. & Garoufallou, E. (2018), 'A Method for Examining Metadata Quality in Open Research Datasets Using the OAI-PMH and SQL Queries: The Case of the Dublin Core "Subject" Element and Suggestions for User-Centred Metadata Annotation Design', *International Journal of Metadata, Semantics and Ontologies* **13**(1), 1—-8. [online] `https://doi.org/10.1504/IJMSO.2018.096444` (Accessed 25 August 2020).

Baro, E. E., Tralagba, E. C. & Ebiagbe, E. J. (2018), 'Knowledge and Use of Self-archiving Options Among Academic Librarians Working in Universities in Africa', *Information and Learning Science* **119**(3/4), 145–160. [online] `https://doi.org/10.1108/ils-01-2018-0003` (Accessed 25 August 2020).

Ben-David, S., Kushilevitz, E. & Mansour, Y. (1997), 'Online Learning versus Offline Learning', *Machine Learning* **29**(1), 45–63. [online] `https://doi.org/10.1023/A:1007465907571` (Accessed 25 August 2020).

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* **3**, 993–1022. [online] `https://www.jmlr.org/papers/v3/blei03a` (Accessed 25 August 2020).

Caragea, C., Wu, J., Gollapalli, S. D. & Giles, C. L. (2016), Document Type Classification in Online Digital Libraries, *in* 'Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence', AAAI'16, AAAI Press, pp. 3997—4002. [online] `https://www.aaai.org/ocs/index.php/IAAI/IAAI16/paper/view/12343/12319` (Accessed 25 August 2020).

Charalampous, A. & Knoth, P. (2017), Classifying Document Types to Enhance Search and Recommendations in Digital Libraries, *in* J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis & I. Karydis, eds, 'Research and Advanced Technology for Digital Libraries', Springer International Publishing, pp. 181–192. [online] `https://doi.org/10.1007/978-3-319-67008-9_15` (Accessed 25 August 2020).

Chilimo, W. (2016), Institutional Repositories: Awareness and Self-archiving Practices of Academic Researchers in Selected Public Universities in Kenya, *in* 'Proceedings of the Fourth CODESRIA Conference on Electronic Publishing, 30 March–1 April 2016, Dakar, Senegal'. [online] `https://codesria.org/IMG/pdf/chilimo.pdf` (Accessed 25 August 2020).

Coulter, N. (1997), 'ACM'S Computing Classification System Reflects Changing Times', *Communications of the ACM* **40**(12), 111–112. [online] `https://doi.org/10.1145/265563.265579` (Accessed 25 August 2020).

Currier, S., Barton, J., O'Beirne, R. & Ryan, B. (2004), 'Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata Creation Process', *ALT-J* **12**(1), 5–20. [online] `https://doi.org/10.1080/0968776042000211494` (Accessed 25 August 2020).

Dhammi, I. & Kumar, S. (2014), 'Medical Subject Headings (MeSH) Terms', *Indian Journal of Orthopaedics* **48**(5), 443–444. [online] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175855` (Accessed 25 August 2020).

Directorate of Research and Graduate Studies (2015), 'Regulations for Postgraduate Training', [online] `https://graduate.unza.zm/images/files/pg_regulations.pdf`. (Accessed 25 August 2020).

DSpace (2018*a*), 'Functional Overview – DSpace 6.x Documentation - LYRASIS Wiki', [online] `https://wiki.lyrasis.org/display/DSDOC6x/Functional+Overview`. (Accessed 25 August 2020).

DSpace (2018*b*), 'Submission User Interface – DSpace 6.x Documentation - LYRASIS Wiki', [online] `https://wiki.lyrasis.org/display/DSDOC6x/Submission+User+Interface`. (Accessed 25 August 2020).

Gries, C., Budden, A., Laney, C., O'Brien, M., Servilla, M., Sheldon, W., Vanderbilt, K. & Vieglais, D. (2018), 'Facilitating and Improving Environmental

Research Data Repository Interoperability', *Data Science Journal*. [online] https://doi.org/10.5334/dsj-2018-022 (Accessed 25 August 2020).

Hafnan, A. P. P. & Mohan, A. (2018), Summary-Based Document Classification, *in* P. K. Sa, S. Bakshi, I. K. Hatzilygeroudis & M. N. Sahoo, eds, 'Recent Findings in Intelligent Computing Techniques', Springer Singapore, Singapore, pp. 153–160. [online] https://doi.org/10.1007/978-981-10-8633-5_16 (Accessed 25 August 2020).

Harnad, S. (2001), 'The Self-archiving Initiative', *Nature* **410**(6832), 1024–1025. [online] http://www.nature.com/articles/35074210 (Accessed 25 August 2020).

Hillmann, D. I. (2008), 'Metadata Quality: From Evaluation to Augmentation', *Cataloging & Classification Quarterly* **46**(1), 65–80. [online] https://doi.org/10.1080/01639370802183008 (Accessed 25 August 2020).

Ioannidis, J. P., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., Contopoulos-Ioannidis, D. G. & Liberopoulos, G. (2007), 'International Ranking Systems for Universities and Institutions: A Critical Appraisal', *BMC Medicine*. [online] http://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-5-30 (Accessed 25 August 2020).

Joblib Developers (2008), 'Joblib: Running Python Functions as Pipeline Jobs', [online] https://joblib.readthedocs.io. (Accessed 25 August 2020).

Katayoon, K. & Abrizah, A. (2010), 'Librarians' Role as Change Agents for Institutional Repositories: A Case of Malaysian Academic Libraries', *Malaysian Journal of Library & Information Science* **15**(3), 121–133. [online] http://eprints.um.edu.my/id/eprint/11898 (Accessed 25 August 2020).

Kim, J. (2010), 'Faculty self-archiving: Motivations and barriers', *Journal of the American Society for Information Science and Technology* **61**(9), 1909–1922. [online] https://doi.org/10.1002/asi.21336 (Accessed 25 August 2020).

Kim, J. (2011), 'Motivations of Faculty Self-archiving in Institutional Repositories', *The Journal of Academic Librarianship* **37**(3), 246–254. [online] https://doi.org/10.1016/j.acalib.2011.02.017 (Accessed 25 August 2020).

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. & Brown, D. (2019), 'Text Classification Algorithms: A Survey', *Information* **10**(4), 150. [online] https://doi.org/10.3390/info10040150 (Accessed 25 August 2020).

Lagoze, C., Van de Sompel, H., Nelson, M. & Warner, S. (2002), 'The Open Archives Initiative Protocol for Metadata Harvesting', [online] http://www.openarchives.org/OAI/openarchivesprotocol.html. (Accessed 25 August 2020).

Lagoze, C., Van de Sompel, H., Nelson, M. L., Warner, S., Sanderson, R. & Johnston, P. (2008), 'Object Re-Use & Exchange: A Resource-Centric Approach', *0804.2273*. [online] http://arxiv.org/abs/0804.2273 (Accessed 25 August 2020).

Lewis, S. (2012), 'SWORD: Facilitating Deposit Scenarios', *D-Lib Magazine*. [online] http://www.dlib.org/dlib/january12/lewis/01lewis.html (Accessed 25 August 2020).

Lynch, C. A. (2003), 'Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age', *portal: Libraries and the Academy* **3**(2), 327–336. [online] https://doi.org/10.1353/pla.2003.0039 (Accessed 25 August 2020).

Mao, Y. & Lu, Z. (2017), 'MeSH Now: Automatic MeSH indexing at PubMed scale via learning to rank', *Journal of Biomedical Semantics* **8**(1), 1–9. [online] https://doi.org/10.1186/s13326-017-0123-3 (Accessed 25 August 2020).

National Library of Medicine (2019*a*), 'Download MEDLINE/PubMed Data', [online] https://www.nlm.nih.gov/databases/download/pubmed_medline.html. (Accessed 25 August 2020).

National Library of Medicine (2019*b*), 'MeSH Record Types', [online] https://www.nlm.nih.gov/mesh/intro_record_types.html. (Accessed 25 August 2020).

Park, J.-R. (2009), 'Metadata Quality in Digital Repositories: A Survey of the Current State of the Art', *Cataloging & Classification Quarterly* **47**(3-4), 213–228. [online] https://doi.org/10.1080/01639370902737240 (Accessed 25 August 2020).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830. [online] http://www.jmlr.org/papers/v12/pedregosa11a.html (Accessed 25 August 2020).

Phillips, M. E., Tarver, H. & Zavalina, O. (2019), 'Using Metadata Record Graphs to Understand Controlled Vocabulary and Keyword Usage for Subject Representation in the UNT Theses and Dissertations Collection', *Cadernos BAD* **2019**(1), 61–76. [online] https://www.bad.pt/publicacoes/index.php/cadernos/article/view/2024 (Accessed 25 August 2020).

Phiri, L. (2018), Research Visibility in the Global South : Towards Increased Online Visibility of Scholarly Research Output in Zambia, *in* 'Proceedings of the 2^{nd} IEEE International Conference in Information and Communication Technologies (ICICT 2018)', Lusaka, Zambia. [online] http://dspace.unza.zm/handle/123456789/5723 (Accessed 25 August 2020).

Phiri, L. (2020), 'lightonphiri/etd_autoclassifier v1.0 (version v1.0)', *Zenodo*. [online] https://doi.org/10.5281/zenodo.3911551 (Accessed 25 August 2020).

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

Rousidis, D., Garoufallou, E., Balatsoukas, P. & Sicilia, M.-A. (2014*a*), Data Quality Issues and Content Analysis for Research Data Repositories: The Case of Dryad, *in* P. Polydoratou & M. Dobreva, eds, 'Proceedings of the 18^{th} International Conference on Electronic Publishing', IOS Press, pp. 49—-58. [online] https://dx.doi.org/10.3233/978-1-61499-409-1-49 (Accessed 25 August 2020).

Rousidis, D., Garoufallou, E., Balatsoukas, P. & Sicilia, M.-A. (2014*b*), 'Metadata for Big Data: A Preliminary Investigation of Metadata Quality Issues in Research Data Repositories', *Inf. Serv. Use* **34**(3–4), 279—-286. [online] https://doi.org/10.3233/ISU-140746 (Accessed 25 August 2020).

Schirrwagen, J., Subirats-Coll, I. & Shearer, K. (2016), 'Coar Resource Types – A Skosified Vocabulary For Open Repositories'. [online] https://doi.org/10.5281/zenodo.55648 (Accessed 25 August 2020).

Sebastiani, F. (2002), 'Machine Learning in Automated Text Categorization', **34**(1), 1–47. [online] https://doi.org/10.1145/505282.505283 (Accessed 25 August 2020).

Steward, S. (2019), 'PDFtk - The PDF Toolkit', [online] https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit. (Accessed 25 August 2020).

Suleman, H. (2011), Interoperability in Digital Libraries, *in* 'E-Publishing and Digital Libraries: Legal and Organizational Issues', IGI Global, chapter 2, pp. 31–47. [online] http://pubs.cs.uct.ac.za/id/eprint/680 (Accessed 25 August 2020).

Suleman, H. (2012), The NDLTD Union Catalog: Issues at a Global Scale, *in* 'Proceedings of the 15^{th} International Symposium on Electronic Theses and Dissertations', Universidad Peruana de Ciencias Aplicadas (UPC). [online] https://repositorioacademico.upc.edu.pe/handle/10757/622568 (Accessed 25 August 2020).

Tani, A., Candela, L. & Castelli, D. (2013), 'Dealing With Metadata Quality: The Legacy of Digital Library Efforts', *Information Processing & Management* **49**(6), 1194–1205. [online] https://doi.org/10.1016/j.ipm.2013.05.003 (Accessed 25 August 2020).

The Networked Digital Library of Theses and Dissertations (2015), 'ETD-MS v1.1: An Interoperability Metadata Standard for Electronic Theses and Dissertations', [online] http://www.ndltd.org/standards/metadata. (Accessed 25 August 2020).

The Pallets Project (2010), 'Flask — The Pallets Project', [online] https://palletsprojects.com/p/flask. (Accessed 25 August 2020).

U.S. National Library of Medicine (2018), 'Overview of Annual Baseline Distribution of MEDLINE/PubMed Data', [online] https://www.nlm.nih.gov/bsd/licensee/baseline.html. (Accessed 25 August 2020).

Webley, L., Chipeperekwa, T. & Suleman, H. (2011), Creating a National Electronic Thesis and Dissertation Portal in South Africa, *in* 'Proceedings of the 14^{th} International Symposium on Electronic Theses and Dissertations', pp. 13–17. [online] http://pubs.cs.uct.ac.za/id/eprint/748 (Accessed 25 August 2020).

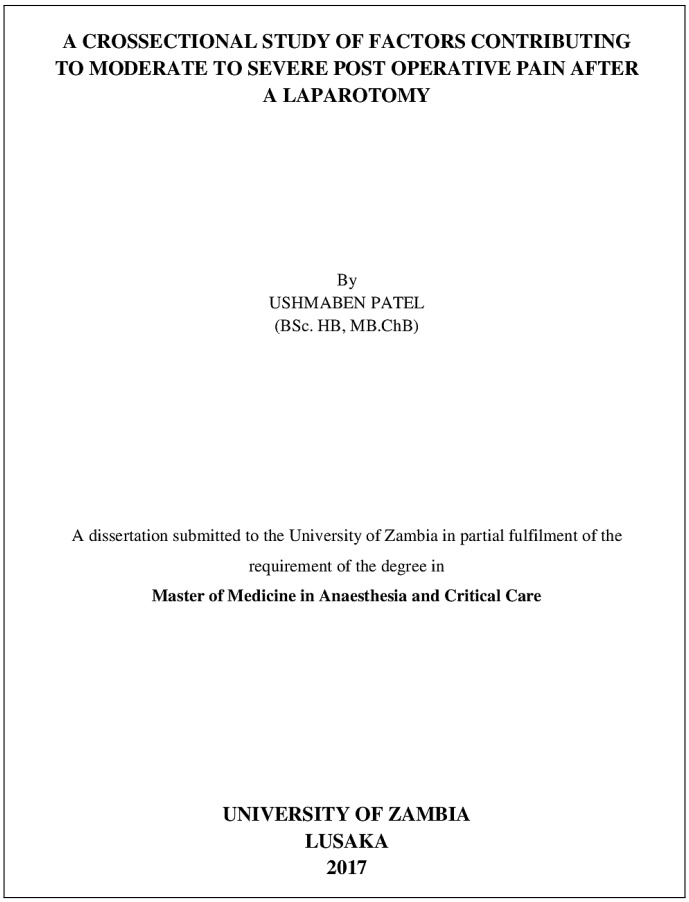
Weibel, S. L., Kunze, J. A., Lagoze, C. & Wolf, M. (1998), 'Dublin Core Metadata for Resource Discovery', [online] http://www.hjp.at/doc/rfc/rfc2413.html. (Accessed 25 August 2020).

Wirth, R. & Hipp, J. (1995), CRISP-DM : Towards a Standard Process Model for Data Mining, *in* 'Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining', pp. 29–39. [online] http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf (Accessed 25 August 2020).

Xia, J. (2007), 'Assessment of Self-archiving in Institutional Repositories: Across Disciplines', *Journal of Academic Librarianship* **33**(6), 647–654. [online] https://doi.org/10.1016/j.acalib.2007.09.020 (Accessed 25 August 2020).

Notes

[1] http://union.ndltd.org/portal
[2] https://oatd.org
[3] https://arxiv.org
[4] https://dl.acm.org
[5] https://www.ncbi.nlm.nih.gov/pubmed
[6] https://jpeg.org
[7] http://www.w3.org/TR/PNG
[8] https://www.pdfa.org/pdfa-facts
[9] http://releases.ubuntu.com/18.04.3



A CROSSECTIONAL STUDY OF FACTORS CONTRIBUTING TO MODERATE TO SEVERE POST OPERATIVE PAIN AFTER A LAPAROTOMY

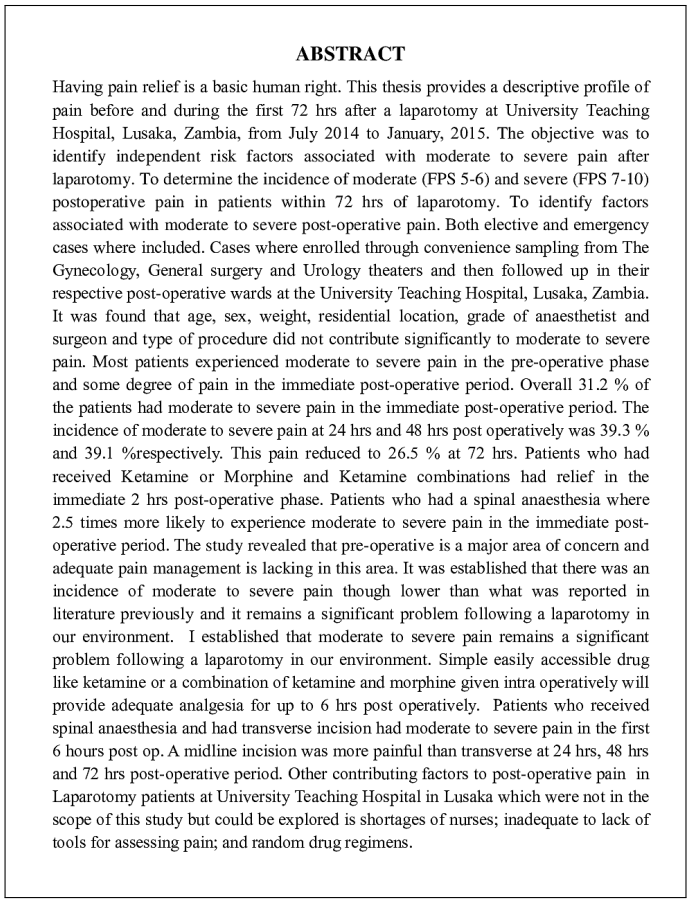
By
USHMABEN PATEL
(BSc. HB, MB.ChB)

A dissertation submitted to the University of Zambia in partial fulfilment of the requirement of the degree in

Master of Medicine in Anaesthesia and Critical Care

UNIVERSITY OF ZAMBIA
LUSAKA
2017

Figure 7: Sample Title Page for ETD Manuscript



ABSTRACT

Having pain relief is a basic human right. This thesis provides a descriptive profile of pain before and during the first 72 hrs after a laparotomy at University Teaching Hospital, Lusaka, Zambia, from July 2014 to January, 2015. The objective was to identify independent risk factors associated with moderate to severe pain after laparotomy. To determine the incidence of moderate (FPS 5-6) and severe (FPS 7-10) postoperative pain in patients within 72 hrs of laparotomy. To identify factors associated with moderate to severe post-operative pain. Both elective and emergency cases where included. Cases where enrolled through convenience sampling from The Gynecology, General surgery and Urology theaters and then followed up in their respective post-operative wards at the University Teaching Hospital, Lusaka, Zambia. It was found that age, sex, weight, residential location, grade of anaesthetist and surgeon and type of procedure did not contribute significantly to moderate to severe pain. Most patients experienced moderate to severe pain in the pre-operative phase and some degree of pain in the immediate post-operative period. Overall 31.2 % of the patients had moderate to severe pain in the immediate post-operative period. The incidence of moderate to severe pain at 24 hrs and 48 hrs post operatively was 39.3 % and 39.1 %respectively. This pain reduced to 26.5 % at 72 hrs. Patients who had received Ketamine or Morphine and Ketamine combinations had relief in the immediate 2 hrs post-operative phase. Patients who had a spinal anaesthesia where 2.5 times more likely to experience moderate to severe pain in the immediate post-operative period. The study revealed that pre-operative is a major area of concern and adequate pain management is lacking in this area. It was established that there was an incidence of moderate to severe pain though lower than what was reported in literature previously and it remains a significant problem following a laparotomy in our environment. I established that moderate to severe pain remains a significant problem following a laparotomy in our environment. Simple easily accessible drug like ketamine or a combination of ketamine and morphine given intra operatively will provide adequate analgesia for up to 6 hrs post operatively. Patients who received spinal anaesthesia and had transverse incision had moderate to severe pain in the first 6 hours post op. A midline incision was more painful than transverse at 24 hrs, 48 hrs and 72 hrs post-operative period. Other contributing factors to post-operative pain in Laparotomy patients at University Teaching Hospital in Lusaka which were not in the scope of this study but could be explored is shortages of nurses; inadequate to lack of tools for assessing pain; and random drug regimens.

Figure 8: Sample Abstract Page for ETD Manuscript

[10] `https://www.xpdfreader.com/pdftotext-man.html`
[11] `https://imagemagick.org/script/convert.php`
[12] `https://github.com/tesseract-ocr`
[13] `https://www.xpdfreader.com/pdfinfo-man.html`
[14] `https://datalab-apis.herokuapp.com/api/collection`
[15] `https://datalab-apis.herokuapp.com/api/type`