Improved Discoverability of Digital Objects in Institutional Repositories Using Controlled Vocabularies

Bertha Chipangila†, Eric Liswaniso†, Andrew Mawila†, Philomena Mwanza†, Daisy Nawila†, Robert M'sendo‡, Mayumbo Nyirenda‡, and Lighton Phiri†
†Department of Library and Information Science, University of Zambia, P.O. Box 32379, Lusaka, Zambia
‡Department of Computer Science, University of Zambia, P.O Box 32379, Lusaka, Zambia
Email: {13000438,15058590,15014576,15018148,15019551,20171520216}@student.unza.zm, mayumbo.nyirenda@cs.unza.zm, lighton.phiri@unza.zm

***Abstract*—Higher Education Institutions (HEIs) utilise Institutional Repositories (IRs) to electronically store and make available scholarly research output produced by faculty staff and students. With the continued increase of scholarly research output produced, accurate and comprehensive association of subject headings to digital objects, during ingestion into IRs is crucial for effective discoverability of the objects and, additionally facilitating the discovery of related content. This paper outlines a case study conducted at an HEI—The University of Zambia—in order to demonstrate the effectiveness of integrating controlled subject vocabularies during the ingestion of digital objects in to IRs. A situational analysis was conducted to understand how subject headings are associated with digital objects and to analyse subject headings associated with already ingested digital objects. In addition, an exploratory study was conducted to determine domain-specific subject headings to be integrated with the IR. Furthermore, a usability study was conducted in order to comparatively determine the usefulness of using controlled vocabularies during the ingestion of digital objects into IRs. Finally, multi-label classification experiments were carried out where digital objects were assigned with more than one class. The results of the study revealed that the majority of digital objects are currently associated with two or less subject headings ($71.2\,\%$), with a significant number of subject headings ($92.1\,\%$) being associated with a single publication. The comparative study suggests that IRs integrated with controlled vocabularies are perceived to be more usable (SUS Score = 68.9) when compared with IRs without controlled vocabularies (SUS Score = 66.2). The effectiveness of the multi-label arXiv subjects classifier demonstrates the viability of integrating automated techniques for subject classification.**

***Keywords*-Controlled Vocabularies; Digital Libraries; Document Classification; Institutional Repositories;**

I. Introduction

Institutional Repositories (IRs) are a crucial part of contemporary Higher Educational Institutions (HEIs) as they provide an avenue for making available scholarly research output produced by faculty staff and students. IRs provide a platform for capturing, preserving and facilitating access to digital work produced by a community [1].

Scholarly research output are typically stored as digital objects within the IRs, with the objects loosely comprising of digital object bitstreams and digital object metadata. Metadata, and more specifically descriptive metadata, is vital for ensuring effective discoverability of the digital objects in IRs.

Fig. 1: A screenshot showcasing sample subjects associated with ingested digital objects in The UNZA's IR.

The University of Zambia (UNZA) has a functional IR with scholarly output consistently deposited into the it, however, there are a number of inconsistencies associated with digital object metadata elements used to describe subject categories related to the objects. Prior work done has identified the lack of use of controlled vocabulary sets as being one of the leading causes of ineffective searching and browsing of scholarly research output in UNZA's IR [2]. In addition to the lack of use of controlled vocabularies, the lack of use of subject specific controlled vocabularies has the potential to make it difficult for end users to search and browse for domain

specific content and related content. These critical anomalies are observable from UNZA's IR: Figure 1 illustrates how the extent to which subjects are inconsistently used, while Figure 2 illustrates how a digital object produced in the Department of Computer Science is associated with non domain-specific subjects.

Fig. 2: A screenshot showcasing subjects associated with a sample Computer Science digital object in UNZA's IR.

This paper presents a study conducted at UNZA to investigate the effectiveness of integrating controlled subject vocabulary sets within UNZA's IR. The study comprised of three phases. First, a situational analysis was conducted to empirically determine the implications of the lack of integration of controlled vocabularies within the repository. In order to understand the potential sources of errors when preparing descriptive metadata, focus group discussions were held with Library staff that administer the IR. Secondly, interview sessions were held with faculty staff in order to identify controlled vocabularies used in their respective domains. Finally, a controlled experiment was designed to empirically determine the usability of IRs integrated with subject controlled vocabularies.

The remainder of this paper is organised as follows: Section II is a synthesis of existing literature related to this work, Section III describes the methodology associated with this work, Section IV presents and discusses the results of this study and, finally, Section V outlines concluding remarks.

II. Related Work

There is a large body of existing literature that has focused on the role of descriptive metadata in facilitating discoverability of digital objects and, the significance of using controlled vocabularies and authority control during ingestion of digital content.

A. Digital Object Descriptive Metadata

IR digital object metadata can be broadly categorised into into the three groups of metadata—administrative metadata, descriptive metadata and structural metadata—proposed by Riley [3]. The metadata is specified as part of an ingestion workflow, with the metadata associated to the digital object externally, as opposed to embedding it within the digital object. While all the three types of metadata are important, descriptive metadata specifically serves the purpose of facilitating the discovery of digital objects through searching and browsing services.

Arms highlights that information discovery is a complex process that can be made effective by referencing descriptive metadata about digital objects stored in repositories [4]. Similarly, Varlamis and Apostolakis emphasise the importance of labelling learning objects stored in learning object repositories in a consistent manner, in order to support indexing and discovery of the content [5]. The importance of labelling is further supported by Currier et al. who state that quality metadata, in particular, enables users to discover and retrieve digital objects in an efficient and effective manner [6].

The external metadata in IRs is encoded using internationally recognised metadata schemes, with Dublin Core [7] being the most widely integrated in popular open source IR software platforms. The digital object metadata is primarily indexed and used to facilitate searching and browsing, however, it is generally possible to activate full-text searching for text-based content.

UNZA's IR is powered by the DSpace open source repository platform. DSpace is capable of processing textual content for full-text searching, in addition to utilising metadata elements during indexing [8]. DSpace uses a default metadata registry that is derived from the 15 Dublin Core metadata elements, with the element values specified during ingestion of digital objects. One of the crucial metadata elements is the "dc.subject" element that specifies the topic associated with the resource

B. Controlled Vocabularies

Controlled vocabularies and authority control are popular techniques that are used to enhance access to bibliographic materials. Harpring defines controlled vocabularies as well-organised words and phrases that are used to index digital content and subsequently facilitate retrieval of the content through searching and browsing [9].

Subject headings are a form of controlled vocabularies that are used to describe topics associated to digital content, making it possible for content related content to be group together. While generic subject headings such as the Library of Congress Subject Headings (LCSH) are widely used, there are other domain specific subject headings, popular with academic databases. For instance, the Medical Subject Headings (MeSH) [10] terms are used in the medical field and the ACM Computing Classification System (CCS) [11] ontology is common used in computing disciplines.

Prior work on subject headings has mostly focused on the effectiveness of subject headings when compared with keywords. In a study aimed at comparing user tags and LCSH Rolla notes that user supplied tags can be used to

enhance subject access but cannot replace the valuable role of controlled vocabularies [12]. This observation supports the results obtained by Lu et al. in a study that suggests that the existence of non-subject-related tags can improve the accessibility of collections [13].

In this work, we empirically determine the implications of sparing use of subject headings and, additionally, identify potential domain-specific subject headings that can be incorporated into IRs. Furthermore, we demonstrate the positive effect subject headings have on the overall usability of IRs.

C. Multi-Label Classification

Motivated by the ever increasingly vast number of digital objects and enhancement in machine learning and technology, multi-label classification has become an extensive studied problem. Multi-label context has in the recent years been researched much because of its application to a wide variety of domains. For example, Konstantions and Kalliris [14] dealt with the problem of automatic detection of emotions in music. Their work established the relation between music and emotion and further looked at multi-labelling mapping of music into emotions. Runzhi et al. used multi-label classification to deal with the problem of multi-disease risk prediction [15]. They constructed a model for prediction of multi-diseases risk relying on the big physical examination data. They acknowledged that in medical diagnosis, a symptom may be associated with various disease types. Chalkidis et al. apply Extreme Multi-Label Text Classification (XMTC) in the legal domain [16]. They employ neural classifiers that outperform the current multi-label state-of-the-art methods, which employ label-wise attention. Boutell et al. focused on video and photography analysis [17]. In semantic scene classification, a picture can be associated to more than on conceptual class such as a sunset and beaches at the same time.

In this work, a multi-label classifier is implemented using an external data source, by taking advantage of transfer learning, and subsequently applied to digital objects associated with the Computer Science field in UNZA's IR, in order to predict appropriate domain-specific subjects.

III. Methodology

The study took a mixed-methods approach involving a situational analysis (See Section III-B), an exploratory study aimed at identifying appropriate subject controlled vocabularies to integrated with the IR (see Section III-C), a usability study aimed at empirically evaluating the effect of controlled vocabularies when integrated with IRs (see Section III-D) and, implementation of a supervised machine learning multi-label classifier (see Section III-E).

A. Datasets

Four datasets were constructed and used to perform the situational analysis, outlined in Section III-B and, additionally, to validate the multi-label classification a model implemented as outlined in Section III-E2. Table I provides a summary of the datasets, with details outlined in Sections III-A1 to III-A4.

TABLE I: Datasets used during experimentation.

Dataset	Objects	Study
UNZA IR	7,440	Situational Analysis
CS@ UCT Archive	995	Situational Analysis · Model Validation
arXiv CoRR	328,011	Situational Analysis
NDLTD Union Catalog	7,296,562	Situational Analysis

1) Dataset #1: arXiv CoRR Dataset: The dataset used for implementing the multi-label classifier was constructed by harvesting Dublin Core [7] encoded metadata records from the arXiv Computing Research Repository (CoRR) [18]. The CoRR specific digital objects were filtered by restricting the harvesting using the OAI-PMH 'SetSpec' verb[1].

General preprocessing operations—removal of punctuations, stemming and stopword removal—were performed on the collected data. However, in addition, non-computing subjects had to be removed from dataset observations, as arXiv CoRR comprises of digital objects tagged with subjects such as Mathematics and Physics.

The constructed dataset comprises of 328,011 digital objects, ingested into CoRR between 2007 and 2021. In addition, the digital objects were tagged with combination of ACM CCS subjects, arXiv subject classes and a combination of the two subject classes. Figure 3 shows the distribution of the subject classes.

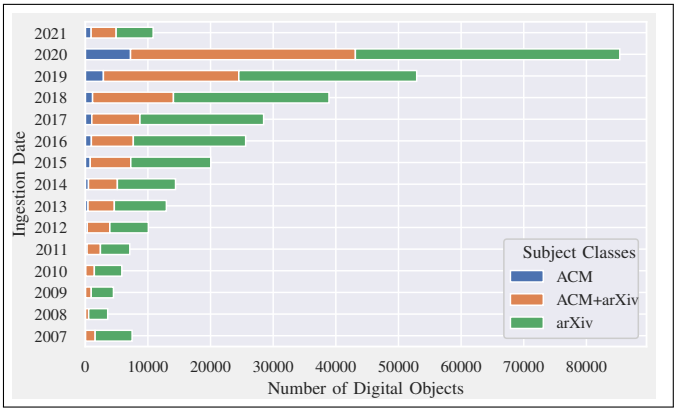

Fig. 3: Distribution of subject tags in the arXiv CoRR dataset.

2) Dataset #2: NDLTD Union Catalog Dataset: A dataset for conducting a situational analysis, outlined in Section III-B, was constructed by harvesting Dublin core encoded metadata records from the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog [19], [20]. 7,404,617 digital object metadata were harvested, with 7,296,562 of them constituting the final dataset, after preprocessing.

3) Dataset #3: UCT CS Document Archive Dataset: A dataset for validating the model was constructed by harvesting Dublin Core encoded metadata records from a Computer Science subject repository (CS@ UCT archive) that is hosted by the Department of Computer Science at The University

[1]http://export.arxiv.org/oai2?verb=ListRecords&metadataPrefix=oai_dc&set=cs

of Cape Town [21]. UNZA's IR has very few Computer Science generated digital objects and as such, it was essential to identify an alternative IR. A total of 1,045 digital object metadata were harvested using the OAI-PMH protocol, with 995 comprising the final dataset, after applying traditional preprocessing operations to remove duplicates, stopwords, punctuations and, additionally, apply stemming. Furthermore, all digital objects with missing titles and abstracts were removed from the dataset.

Faculty and postgraduate students self-archive digital objects into the CS@ UCT archive. More importantly, however, the 2012 ACM CCS concepts are used as the primary controlled vocabulary set. Owing to the fact that the arXiv CoRR dataset described in Section III-A1 uses the 1998 ACM CCS concepts, the CS@ UCT archive dataset was used to compare the distribution of subject classes in order to demonstrate the effectiveness of the multi-label classification model implemented, as described in Section III-E2.

4) Dataset #4: UNZA IR Dataset: A dataset for conducting a situational analysis, outlined in Section III-B, was constructed by harvesting Dublin Core encoded metadata records from UNZA's IR. The 'identifier' and 'subject' Dublin Core elements were used to assess the distribution of manually subject tags. A total of 5,440 metadata records were harvested, with 4,802 constituting the final dataset after basic preprocessing.

B. Situational Analysis

1) Empirical Analysis of UNZA IR: Digital objects ingested into UNZA's IR can be broadly classified into two groups: faculty produced scholarly output—pre-print and post-print versions of peer-reviewed publications—and student produced scholarly output—Electronic Theses and Dissertations.

Ingestion of faculty produced scholarly output is not into UNZA's repository is not consistent, in part due to the lack of availability of an IR policy. However, ETDs are routinely ingested into the IR.

Digital object structural and Dublin Core encoded descriptive metadata, associated with ETDs, were thus harvested from UNZA's IR using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [22]. Specifically, the ListRecord verb[2] was used in conjunction with the ListSets verb[3].

The structural metadata—Lines 7–8 in Listing 1—was required to identify the subject domains associated with the ETDs, while the descriptive metadata was required to identify subjects—Lines 17–19 in Listing 1— associated with the ETDs when ingested into the IR.

2) Empirical Analysis of Selected Portals: While the focus of this study was on UNZA's IR, in order to demonstrate the severity of the problem, a basic analysis of two additional scholarly portals was conducted. A reasonably sized Computer Science subject repository, hosted by the Department of Computer Science at The University of Cape Town was analysed in order to highlight the lack of comprehensive usage of subject controlled vocabularies. In addition, a large scale portal, the NDLTD Union Catalog was analysed in order to demonstrate the implications of lack of comprehensive use of subject controlled vocabularies on a global scale.

3) Digital Object Ingestion Workflow: In order to understand how digital objects are ingested into UNZA's IR, a focus group discussion was conducted with two Library members of staff that are tasked with preparing digital object metadata and ingestion of digital objects into the repository.

Listing 1: A sample ETD metadata record harvesting using the OAI-PMH protocol ListRecords verb.

```
1  <record>
2   <header>
3    <identifier>
4     oai:dspace.unza.zm:123456789/6413
5    </identifier>
6    <datestamp>2020-09-21T10:38:06Z</datestamp>
7    <setSpec>com_123456789_18</setSpec>
8    <setSpec>col_123456789_84</setSpec>
9   </header>
10  <metadata>
11   <oai_dc:dc>
12    <dc:title>
13     Automation of the grain purchasing Process for'
14     Zambias food reserve Agency
15    </dc:title>
16    <dc:creator>Simukanga, Alinani</dc:creator>
17    <dc:subject>Agricultural informatics</dc:subject>
18    <dc:subject>Agriculture-Data processing.</dc:subject>
19    <dc:subject>Agricultural innovations.</dc:subject>
20    <dc:description>
21     [...]
22     The aim of this work is to automate the processes of
23     FRA, FISP and the Cooperatives Society operate, with
24     a specific focus on the farmer registry and the grain
25     marketing process.
26     [...]
27    </dc:description>
28    <dc:date>2020-09-21T10:38:03Z</dc:date>
29    <dc:date>2020-09-21T10:38:03Z</dc:date>
30    <dc:date>2019</dc:date>
31    <dc:type>Thesis</dc:type>
32    <dc:identifier>
33     http://dspace.unza.zm/handle/123456789/6413
34    </dc:identifier>
35    <dc:language>en</dc:language>
36    <dc:format>application/pdf</dc:format>
37    <dc:publisher>University of Zambia</dc:publisher>
38   </oai_dc:dc>
39  </metadata>
40 </record>
```

C. Identification of Appropriate Controlled Vocabularies

Seven faculty staff, from UNZA, were purposively sampled in order to elicit information about possible subject controlled vocabularies associated with the various disciplines at UNZA. Semi-structured face-to-face interview sessions were then conducted with each of the individual faculty staff. The interview sessions were recorded and, additionally, notes taken during the sessions.

D. Usability of IRs Integrated With Controlled Vocabularies

In order to empirically demonstrate the usability effect of integrating IRs with subject controlled vocabularies, a controlled experiment was conducted by comparing a baseline IR setup without a controlled vocabularies and a control IR integrated

[2]http://dspace.unza.zm/oai/request?verb=ListRecords&metadataPrefix=oai_dc

[3]http://dspace.unza.zm/oai/request?verb=ListSets

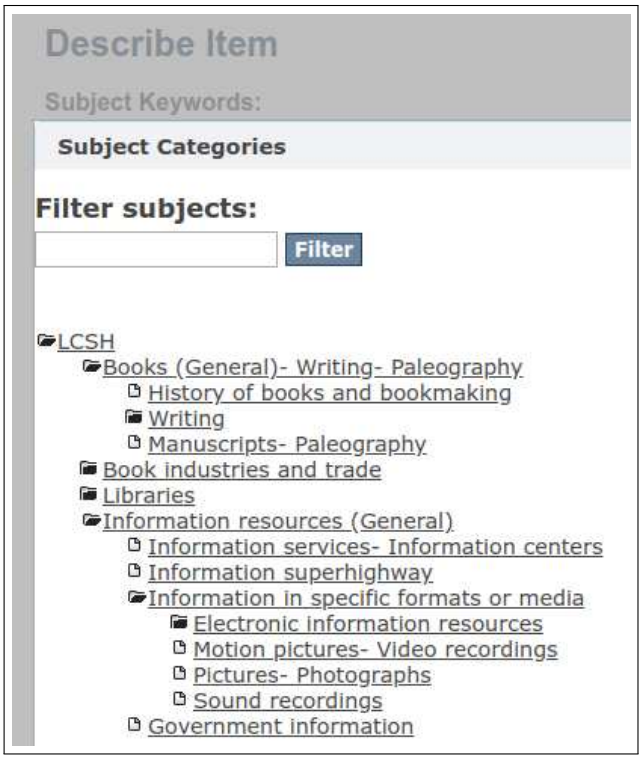

Fig. 4: A screenshot showing the integration of LCSH vocabulary in the intervention IR used for experimentation.

with the LCSH vocabulary set, as shown in Figure 4. Both repositories were setup using DSpace 6.x, with the intervention IR integrated with controlled vocabularies using hierarchical controlled LCSH vocabularies [23].

1) Prototype Institutional Repository Platforms: Two prototype DSpace-powered IRs were installed, setup and configured to be used to conduct the experiment. A baseline IR was setup without integrating it with controlled vocabularies, while the control IR was integrated with LCSH controlled vocabulary set.

2) Experimental Design: A within subject experiment was designed, using random experiment blocks. 50 undergraduate students were randomly sampled from the Dept. Library and Information Science at UNZA, to participate in the study. Each of the 50 participants ingested a digital object, using the two prototype IR that were setup. In each instance, participants filled out a System Usability Score (SUS) questionnaire upon successful ingestion of the digital object.

E. Multi-Label Subject Classifier

A multi-label classifier was implemented by taking advantage of transfer learning, with the model implemented using data from an external repository and, subsequently applied to new observations in UNZA's IR.

1) Data Preparation: As mentioned in Section III-A1, the arXiv CoRR dataset was used during experimentation of the multi-label classification model. The data attributes of the arXiv dataset were used as follows:

- Identifier—A unique identifier for uniquely identifying each of the arXiv digital objects harvested.
- Title—The arXiv digital object publication title, used to extract text input features.
- Description—The arXiv digital object abstract, used to extract text input features.
- Subject—The arXiv-specific digital object subjects [24] and 1998 ACM Computing Classification System (CCS) concepts [25], used as labels.

2) Model Implementation: The model features were extracted from the digital object publication titles and abstracts, with subsequent transformation of the input features done using CountVectorizer and TFIDFVectorizer. The model was implemented using the scikit-multilearn Python library [26]. Binary Relevance and Classifier Chains approaches to multi-label classification were used, in conjunction with estimators—Random Forest and Naive Bayes (Multinomial)—popularly used for text classification. Section IV-D discusses experimental results conducted to experimentally evaluate the effectiveness of the modal features and transformations used.

3) Experimental Design: All experiments were performed 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[4].

Training and testing datasets were created using the hold-out method built within the scikit-multilearn Python library, with 70 % of each dataset used for training and the remaining 30 % for testing.

It is essential to take in account multiple and contrasting metrics measures because of the additional degree of freedom that multi-label introduces and as such, the metrics used to measure the performance for multi-label classification are usually different from those used in binary and multi-class problems. Traditional multi-label classification metrics cited in literature [27]—F1 score, Jaccard Score Similarities and Hamming Loss metrics—were used to evaluate the model.

In order to determine the combination of factors that yield the best results, experimentation involved varying the following aspects:

- Input features—Title, Abstract and a combination of the two: Title+Abstract
- Text transformation techniques—CounterVectorizer and TFIDFVectorizer—used on input features and their corresponding parameters
- Multi-label classification approaches—Binary Relevance and Classifier Chains
- Estimators—Random Forest and Naive Bayes (Multinomial)

Experimentation involved measuring evaluation metrics by varying the experiment factors and aspects mention above. In addition, a validation exercise was conducted, that involved comparing the distribution manually assigned 2012 ACM CCS concepts with the 1998 ACM CCS concepts predicted by

[4]http://releases.ubuntu.com/18.04.3

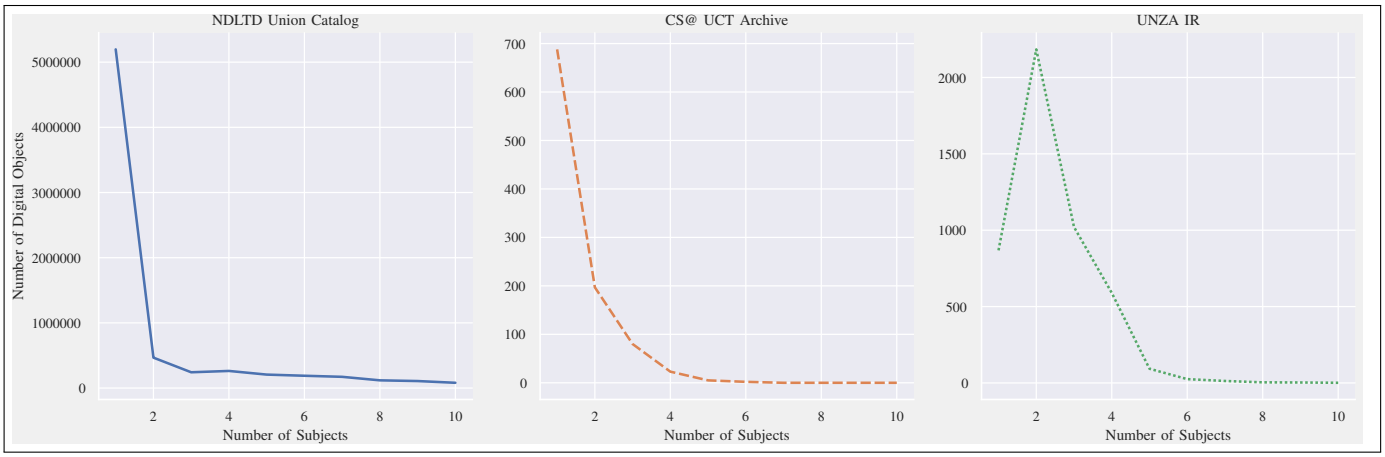

Fig. 5: Digital objects in most digital libraries are typically associated with very few subject classes.

the model, in the CS@ UCT archive dataset described in Section III-A3. Section IV-D presents and discusses the results.

IV. Results and Discussion

A. Situational Analysis

1) Analysis 1. Metadata Preparation Workflow: The focus group session was conducted with two Library staff—the IR Manager and his assistant. The Library staff highlighted that as part of the metadata preparation process, digital objects to are catalogued and subject headings copied from an online public access catalog[5]. This process presents a number of challenges as it is time consuming and error prone. Integration of the IR with appropriate subject headings would not only help address these challenges, but also ensure effective self-archiving [28] of digital objects into the repository.

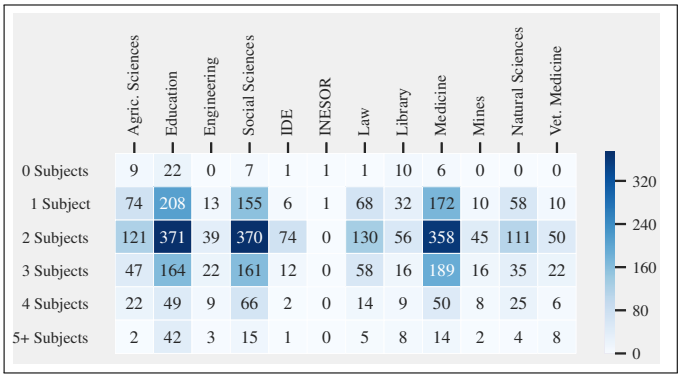

Fig. 6: A heatmap showing the average number of subject/topic specified for scholarly publications for the various domains at UNZA.

2) Analysis 2. Subject Headings for UNZA IR: 5,438 digital object metadata were harvested from UNZA's IR. Scholarly output associated with the 13 faculties at UNZA were then filtered, resulting in a total of 3,638 metadata records used in this analysis. The digital object metadata records were analysed to determine the usage of subject headings.

Of the 3,638 digital objects analysed, 22.2 % were assigned a single subject heading, 47.4 % were assigned two subject headings, 24.4 % were assigned three subject headings, 7.1 % were assigned four subject headings and 2.9 % were assigned more than five subject headings. Figure 6 shows a heatmap of number of subjects assigned to scholarly research output. In the heatmap, it is evident that a significant proportion of scholarly publications, in each of the faculties, are tagged with two subjects. While the number of subjects used to classify a publication is dependent on how the contents of the publication, interviews conducted with Library staff and UNZA revealed that an internal policy requires that digital objects be associated with at least two or three subject headings. However, in an ideal case, it is desirable to associate a publication with more tags to facilitate effective discoverability of related content.

A total of 7,244 subject headings were associated with the data analysed. Of the total subject headings, 92.1 % were associated to a single publication, 6.1 % to two publications, 0.9 % to three publications, 0.1 % to four publications and 0.5 % to more than five publications. Figure 7 shows a heatmap of subject usage patterns for scholarly publications by faculty. The heatmap showcases the frequency of usage of subject headings. For instance, 1,402 subjects have been assigned to only a single publication for content ingested into "Medicine" collections, whereas only 10 subjects are associated with five or more publications. The chart indicates that lack of use of subject controlled vocabularies due to the significantly large proportion of subjects being associated with a single publication. The sparing use of subjects is also shown in Figure 1.

3) Analysis 3. Subject Class Distribution in Portals: Figure 5 show the distribution of subject classes in UNZA's IR, the CS@ UCT Archive and the NDLTD Union Catalog. A common characteristic of the three portals is that most of the digital objects are associated with less than two subjects.

[5]http://koha.unza.zm:4480

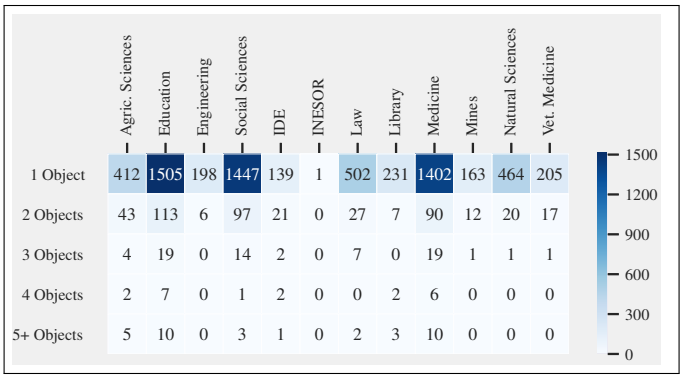

Fig. 7: A heatmap showing the number of subjects/topics associated with different thresholds of scholarly publications for the various domains at UNZA.

While the distribution is especially problematic for portals like UNZA's IR, largely due to low self-archiving practices, the picture is equally as bad for portals such as CS@ UCT Archive, where self-archiving is practised significantly; this is largely due to the fact that authors that self-archive tend not to comprehensively provide relevant subject classes to their publications. Large scale downstream services such as the NDLTD Union Catalog have worse off distributions because the content archiving in such portals is harvested from IRs that have problematic self-archiving practices.

The distributions shown in Section IV-A3 support the premise of this paper: the problem with subject classes is best addressed at the source. Incidentally, it is possible to introduce interventions that can be applied to downstream services, however, it is more effective to work with source portals.

B. Domain-Specific Subject Headings

Seven faculty staff were interviewed in order to elicit subject headings used in their various disciplines. Table II shows a summary of the major outcomes from the interview sessions. Most of the interviewees were familiar with the concept of controlled vocabularies, however, only a few were knowledgeable about the specific subject headings used in their respective domains.

While the majority of faculty staff are unaware of subject headings used in their disciplines, a question included in the interview guide required that they specify popular academic databases used in their domains. The academic databases specified can be used as a basis for identify appropriate subject headings. For instance, the the widely used ACM Computing Classification System [11] could be used to generate subject headings for scholarly research output produced by computing oriented faculties and/or departments. Using academic databases as a basis for adopting subject headings could also potentially enhance the interoperability of IRs with external downstream services that automatically harvest IR metadata.

TABLE II: Summary of results from interviews conducted as part of an exploratory study to understand subject headings used in various domains at UNZA.

Participant	Academic Databases	Subjects
FS–1	PubMed · Science Direct · Google Scholar · Mendeley	MeSH
FS–2	SCOPUS · ERIC · SCINAPSE · EBSCO HOST · PROQUEST	Not aware
FS–3	Academia.edu · Zambia Library Journals · Google Scholar · UNZA IR	Not aware
FS–4	IEEE · ELSEVIER	Not aware
FS–5	ResearchGate · Google Scholar · Academia.edu	None
FS–6	Academia.edu · Mendeley · ResearchGate · JSTOR · Google Scholar	SEARS List
FS–7	IEEE · Explorer · ResearchGate	None

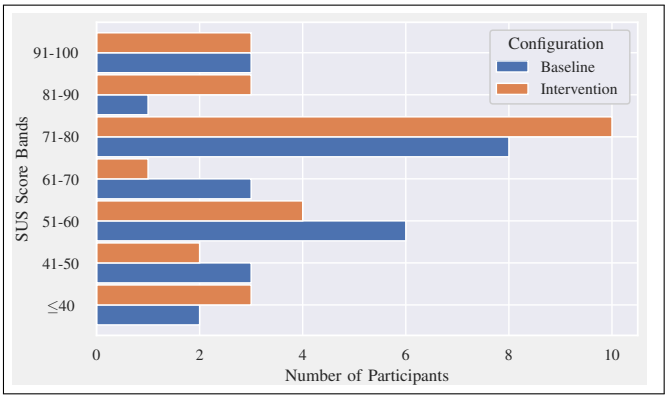

Fig. 8: SUS acceptability and adjective rating scores [29] for the baseline and intervention IRs.

C. Comparative Usability Study

1) System Usability Scale Scores: The SUS scores corresponding to responses from each participants were computed for each of the two IR platforms: baseline and intervention. The SUS scores were calculated using the standard method of that takes into account all of the 10 SUS questionnaire items [30].

The average SUS scores for the baseline and intervention IRs were 66.2 and 68.9, respectively. While both SUS scores are rated "OK" on the acceptability and adjective rating score [29], the average SUS score for the intervention is noticeably higher, as shown in Figure 8. The differences in the SUS scores is further supported by the positive responses associated with the intervention IR, outlined in Section IV-C2.

However, a paired t-test indicates no significant difference in the mean scores ($p = 0.82$). Furthermore, Factorial ANOVA tests conducted to determine effects of demographic factors also suggest no significant main effect as a result of "Prior Knowledge of Controlled Vocabularies" ($F_{1,48} = 0.041$, $p = 0.84$), "Experience With ICTs" ($F_{2,46} = 1.13$, $p = 0.33$), "Participants' Year of Study" ($F_{2,46} = 0.77$, $p = 0.47$) and "Gender" ($F_{1,48} = 1.61$, $p = 0.21$),

2) Participants' Comments: Participants were also required to provide open ended comments, relative to their experiences using the two platforms. The vast majority of the comments were related to the use of controlled vocabularies and, for the most part, positive.

> *"It was easy to work around the repository with subject controlled vocabulary."* [Participant #6]
>
> *"The second method is more easier to work with"* [Participant #14]
>
> *"It was easy because you have to just click and the keywords will be provided which is less time consuming"* [Participant #18]
>
> *"The arrangement is well organised and kind of easy to use"* [Participant #22]
>
> *"I did not like typing in the subject keywords."* [Participant #26]

The participants' comments, in part, help explain the higher SUS mean score for the intervention IR, outlined in Section IV-C1.

TABLE III: Model versus Manual generated subjects.

Digital Object Title	
Automation of the grain purchasing Process for Zambia's food reserve Agency	
Digital Object Abstract	
Issues of food security, post-harvest losses, lack of a national farmer database and proper grain inventory system have plagued the Ministry of Agriculture for years. The lack of requisite tools has made the management of the sector a difficult task. This has seen an increase in the number of ghost farmers benefiting from the Farmer Input Support Programme (FISP). The aim of this work is to automate the processes of FRA, FISP and the Cooperatives Society operate, with a specific focus on the farmer registry and the grain marketing process. The objectives are as follows: Map the current business processes of FISP and FRA; Develop a model of objective 1 using cloud and mobile computing technologies; Develop a system prototype that integrates farmers spatial data and mobile computing based on the model in objective 2; and integrate multi-factor authentication into the prototype in objective 3. To meet objective 1, a baseline study was conducted at the FRA depots in Chongwe and Mumbwa. The information gathered from this and from various documents provided informed the development of the model specified in objective 2. Various web technologies such as PHP, Java and PostgreSQL were employed to achieve objective 3. Multi-factor authentication was implemented as an added security feature when interfacing with the mobile application for the final objective.	
Digital Object Subjects	
Manual Subjects	**Model Generated Subjects**
Agricultural informatics · Agriculture–Data processing · Agricultural innovations	C.2.4 · Computer Science - Artificial Intelligence · Computer Science - Computation and Language · Computer Science - General Literature · Computer Science - Human-Computer Interaction · D.2.11 · F.1.1 · H.3.4 · H.3.5 · H.5.2

D. ArXiv CoRR Subject Classification

Table III illustrates the results obtained when the model is applied to a sample digital object in UNZA's IR. The "Manual Subjects" column has subjects manually prepared when the digital object was being ingested into the IR, while the "Model Generated Subjects" column represents the subjects predicted by the model. The ACM CCS concepts are presented with their code for brevity, while the arXiv subjects are present with their textual descriptions.

Table IV depicts the summary of performance results in terms of F1-score, hamming loss and Jaccard Similarities accuracy for two multi-label classification approaches used, including the estimators used and data transformation technique.

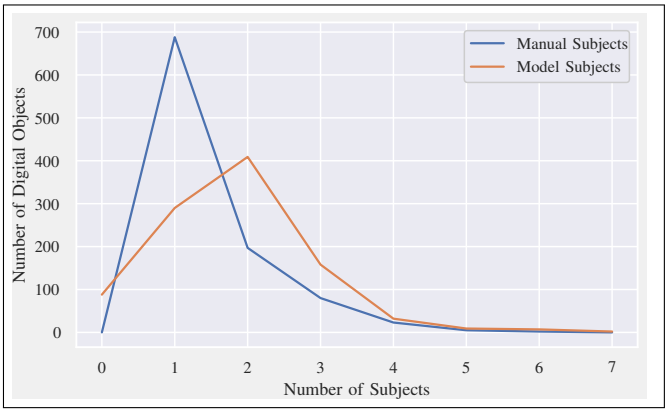

Fig. 9: Distribution of Model and Manual generated subjects.

1) Analysis 1. Transfer Learning: The implemented model was applied to digital objects in UNZA's IR, with very promising results. First, the predicted labels are specific to the domain in question: Computer Science, as opposed to the previous subjects that are manually determined by staff that ingest digital objects into the IR. More significantly, however, randomly inspected objects were noted to have been automatically associated with relevant subjects. Table III shows a comparison of manual subjects previously associated to the sample object, also shown in Figure 2, and subjects automatically predicted by the model. Only three (3) non subject-specific subjects were annually assigned to the digital object, while a total of six ACM CCS subjects and four arXiv subjects were automatically predicted using the model.

Figure 9 shows a subject classes distribution of manually assigned subjects and model generated subjects in the CS@ UCT archive. It is evident from the plot that a significant proportion of digital objects only have a single subject associated with them, a common occurrence in IRs that implement self-archiving.

The automatic prediction of subjects has the obvious benefit of ensuring that a consistent subset of domain-specific subjects are associated with related digital objects. Furthermore, this technique reduces the human-centric manual processes involved in when associating metadata to digital objects, drastically reducing the time spent preparing metadata and potential errors introduced when preparing metadata.

2) Analysis 2. Input Features: As earlier mentioned in Section III-E, three input features were used during experimentation 'Title', 'Abstract' and 'Title+Abstract'. Expectantly,

TABLE IV: Experimental results for arXiv subject classes multi-label classification model.

Binary Relevance										
		Title			Abstract			Title + Abstract		
		F1 Score	Hamming Loss	Jaccard Score	F1 Score	Hamming Loss	Jaccard Score	F1 Score	Hamming Loss	Jaccard Score
MultinomialNB	TF	0.305	0.006	0.192	0.214	0.037	0.207	0.203	0.041	0.196
	TF-IDF	0.236	0.006	0.148	0.398	0.005	0.271	0.420	0.005	0.290
RandomForest	TF	0.317	0.006	0.211	0.416	0.005	0.292	0.430	0.005	0.305
	TF-IDF	0.314	0.006	0.210	0.418	0.005	0.295	0.435	0.005	0.310
SGDClassifier	TF	0.279	0.006	0.18	0.515	0.005	0.390	0.526	0.005	0.407
	TF-IDF	0.282	0.006	0.183	0.476	0.005	0.351	0.496	0.005	0.369
Classifier Chains										
		Title			Abstract			Title + Abstract		
		F1 Score	Hamming Loss	Jaccard Score	F1 Score	Hamming Loss	Jaccard Score	F1 Score	Hamming Loss	Jaccard Score
MultinomialNB	TF	0.060	0.055	0.190	0.030	0.338	0.130	0.030	0.347	0.123
	TF-IDF	0.090	0.027	0.177	0.086	0.053	0.282	0.087	0.055	0.294
RandomForest	TF	0.287	0.009	0.238	0.428	0.005	0.305	0.441	0.005	0.318
	TF-IDF	0.289	0.009	0.239	0.424	0.005	0.301	0.444	0.005	0.320
SGDClassifier	TF	0.312	0.006	0.216	0.527	0.005	0.420	0.520	0.006	0.414
	TF-IDF	0.310	0.006	0.214	0.523	0.005	0.413	**0.540**	0.005	**0.431**
One-Versus-Rest										
		Title			Abstract			Title + Abstract		
		F1 Score	Hamming Loss	Jaccard Score	F1 Score	Hamming Loss	Jaccard Score	F1 Score	Hamming Loss	Jaccard Score
MultinomialNB	TF	0.305	0.006	0.192	0.214	0.037	0.207	0.203	0.041	0.196
	TF-IDF	0.236	0.006	0.148	0.398	0.005	0.271	0.420	0.005	0.290
RandomForest	TF	0.317	0.006	0.212	0.414	0.005	0.291	0.435	0.005	0.310
	TF-IDF	0.315	0.006	0.210	0.414	0.005	0.290	0.432	0.005	0.306
SGDClassifier	TF	0.279	0.006	0.180	0.488	0.005	0.365	0.520	0.005	0.400
	TF-IDF	0.282	0.006	0.183	0.479	0.005	0.354	0.497	0.005	0.371

using a combination of titles and abstracts—Title+Abstract—results in more effective models than using the Title or Abstract features in isolation. It was also noticed that overall transforming the input features using TFIDFVectorizer result in better performing models than using CounterVectorizer. This is the case for the best performing approach and estimator, SGDClassifier using Classifier Chains, where the F1 Score was 0.540 and the Jaccard Similarities Score was 0.431. Incidentally, this is also the case for most of the other approaches and estimators

The 'Title+Abstract' feature Expectantly results in better performance seeing as combining two features results in a more enriched feature. New enriched feature-sets can potentially be created by augmenting metadata elements such as 'Keywords', which are sometimes provided alongside traditional ones like 'Title' and 'Abstract'.

3) Analysis 3. Approach and Estimators: Of the three multi-label approaches used, Classifier Chains yielded the best results, with an F1 score of 0.540; Hamming Loss value of 0.005 and Jaccard Similarities Score of 0.431. The next best performing approach was One-Versus-Rest, using SGDClassifier, and finally, One-Versus-Rest using SGDClassifier. In all these instances, the 'Title+Abstract' yielded the best result.

With the F1 Score results obtained, it makes logical sense that IR interfaces incorporate the automatic generation of subject classes in such a manner that they are complemented with human effort. For instance, an end-user can be presented with an interface that enable them to add and/or remove subject classes automatically generated.

V. Conclusion

This paper outlines a case study conducted to investigate the implications of integrating subject controlled vocabularies in IRs. The case study was conducted in three phases. First, a situational analysis—described in Section III-B—was conducted to understand how digital objects are tagged with subject headings. Secondly, an exploratory study—outlined in Section III-C—was conducted to determine domain specific subject headings for different faculties at UNZA. In addition, a usability study—outlined in Section III-D—was conducted to ascertain the impact on usability of IRs integrated with controlled subject vocabularies. Finally, a multi-label classification model for predicting ACM CCS and arXiv subject classes was presented. Experimental results of the classification model illustrate the potential effectiveness of automatically generating domain specific subjects.

Integrating IRs with subject controlled vocabularies has the benefit ensuring that IRs are usable and effective. More significantly, though, the digital objects are certain to be tagged with correct and comprehensive subject headings. The semi-automatic metadata generation approach proposed, where traditional human approaches are augmented with an automated approach align with techniques suggested by Tani et al. for addressing metadata quality issues [31].

The systematic process presented in this paper has the potential of making self-archiving more effective, since IRs

would be integrated with pre-existing subject headings. This would ultimately complement machine learning techniques presented in prior work [32], further making the ingestion of digital objects into IRs more effective, less error and more comprehensive Beyond IRs, however, the automatic generation of subject classes can be applied to large scale portals such as the NDLTD Union Catalog [19], [20] that have been noted to experience metadata quality issues [33].

As part of future and on-going work, models are being implemented for other domains at UNZA and, additionally, there are plans to apply this technique on large-scale datasets such as the NDLTD Union Catalog [19].

REFERENCES

[1] N. F. Foster and S. Gibbons, "Understanding faculty to improve content recruitment for institutional repositories," *D-Lib Magazine*, vol. 11, no. 1, 2005.

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

[3] J. Riley, *Understanding Metadata: What is Metadata, and what is it For?* National Information Standards Organization (NISO), 2017.

[4] W. Y. Arms, "Information retrieval and descriptive metadata," in *Digital Libraries*. MIT press, 2001, ch. 10.

[5] I. Varlamis and I. Apostolakis, "The Present and Future of Standards for E-Learning Technologies," *Interdisciplinary Journal of e-Skills and Lifelong Learning*, vol. 2, pp. 59–76, 2006.

[6] C. Sarah, B. Jane, O. Rónán, and R. Ben, "Quality assurance for digital learning object repositories: issues for the metadata creation process," *ALT-J*, vol. 12, no. 1, pp. 5–20, 2004.

[7] S. L. Weibel, J. A. Kunze, C. Lagoze, and M. Wolf, "Dublin Core Metadata for Resource Discovery," 1998. [Online]. Available: http://www.hjp.at/doc/rfc/rfc2413.html

[8] DSpace, "Functional Overview - DSpace 6.x Documentation," 2018. [Online]. Available: https://wiki.lyrasis.org/display/DSDOC6x/Functional+Overview

[9] P. Harpring, "What are controlled vocabularies?" in *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Publications, 2010, ch. 2.

[10] I. Dhammi and S. Kumar, "Medical subject headings (MeSH) terms," *Indian Journal of Orthopaedics*, vol. 48, no. 5, p. 443, 2014.

[11] Association for Computing Machinery, "ACM Computing Classification System," 2012. [Online]. Available: https://dl.acm.org/ccs

[12] P. J. Rolla, "User Tags versus Subject Headings," *Library Resources & Technical Services*, vol. 53, no. 3, pp. 174–184, 2013.

[13] C. Lu, J. R. Park, and X. Hu, "User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings," *Journal of Information Science*, vol. 36, no. 6, pp. 763–779, 2010.

[14] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," *ISMIR 2008 - 9th International Conference on Music Information Retrieval*, pp. 325–330, 2008.

[15] R. Li, W. Liu, Y. Lin, H. Zhao, and C. Zhang, "An Ensemble Multi-label Classification for Disease Risk Prediction," *Journal of Healthcare Engineering*, vol. 2017, 2017.

[16] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in eu legislation," in *Proceedings of the Natural Legal Language Processing Workshop*, 2019, pp. 78–87.

[17] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[18] Cornell University, "Welcome to the Computing Research Repository (CoRR)," 2019. [Online]. Available: https://arxiv.org/corr

[19] The Networked Digital Library of Theses and Dissertations, "NDLTD Union Archive of ETD Metadata," [online] http://union.ndltd.org/portal, (Accessed April 1, 2021).

[20] ——, "Global ETD Search," [online] http://search.ndltd.org, (Accessed April 1, 2021).

[21] Department of Computer Science, School of IT, University of Cape Town, "UCT Computer Science Research Document Archive," [online] https://pubs.cs.uct.ac.za, (Accessed April 1, 2020).

[22] C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner, "The Open Archives Initiative Protocol for Metadata Harvesting," 2002. [Online]. Available: http://www.openarchives.org/OAI/openarchivesprotocol.html

[23] DSpace. (2015) Authority Control of Metadata Values - DSpace 6.x Documentation. [Online]. Available: https://wiki.lyrasis.org/display/DSDOC6x/Authority+Control+of+Metadata+Values

[24] Cornell University, "Computer Science Subject Areas and Moderators," 2020. [Online]. Available: https://arxiv.org/corr/subjectclasses

[25] Association for Computing Machinery, "The ACM Computing Classification System (1998)," 2007. [Online]. Available: https://www.acm.org/publications/computing-classification-system/1998/ccs98

[26] P. Szymanski and T. Kajdanowicz, "Scikit-multilearn: a scikit-based python environment for performing multi-label classification," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 209–230, 2019.

[27] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[28] J. Kim, "Faculty self-archiving: Motivations and barriers," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 9, pp. 1909–1922, sep 2010. [Online]. Available: http://doi.wiley.com/10.1002/asi.21336

[29] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," vol. 4, no. 3, pp. 114–123, May 2009. [Online]. Available: http://www.usabilityprofessionals.org/upa_publications/jus/2009may/JUS_Bangor_May2009.pdf

[30] J. Brooke, *SUS – A quick and dirty usability scale*. London: Taylor & Francis, 1996, ch. 21, pp. 189–195. [Online]. Available: http://www.usabilitynet.org/trump/documents/Suschapt.doc

[31] A. Tani, L. Candela, and D. Castelli, "Dealing with metadata quality: The legacy of digital library efforts," *Information Processing & Management*, vol. 49, no. 6, pp. 1194 – 1205, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457313000526

[32] L. Phiri, "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, 2020.

[33] H. Suleman, "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), 2012, [online] https://repositorioacademico.upc.edu.pe/handle/10757/622568 (Accessed April 1, 2021).