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 · Lighton Phiri

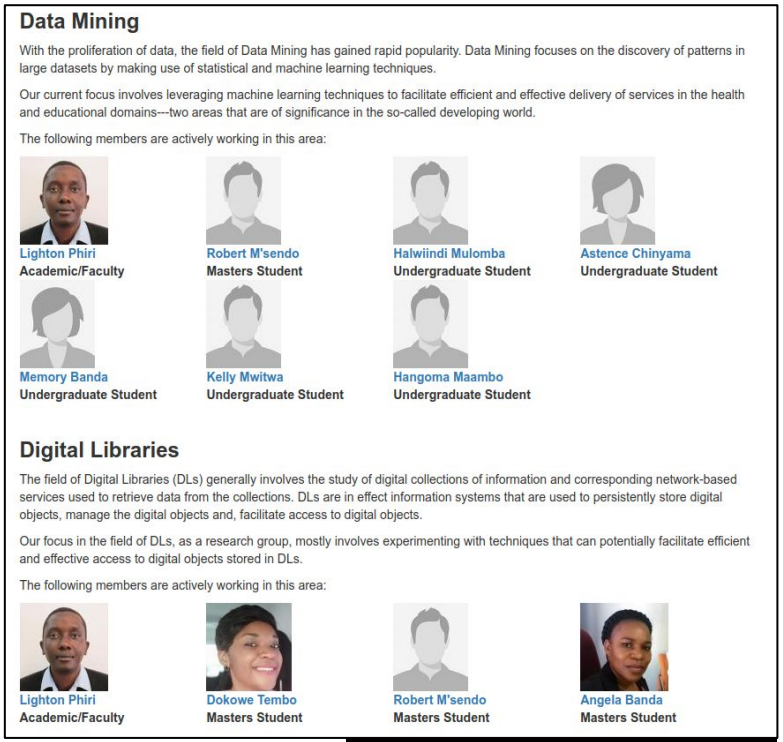
University of Zambia
Lusaka, ZAMBIA

ACM/IEEE Joint Conference on Digital Libraries (JCDL 2021)
September 27–30, 2021

About Us (1/2)

About Us (2/2)

- **The DataLab research group at The University of Zambia is composed of faculty staff and students—undergraduate and postgraduate—working in three main areas**
  - Data Mining
  - Digital Libraries
  - Technology-Enhanced Learning



Data Mining

With the proliferation of data, the field of Data Mining has gained rapid popularity. Data Mining focuses on the discovery of patterns in large datasets by making use of statistical and machine learning techniques.

Our current focus involves leveraging machine learning techniques to facilitate efficient and effective delivery of services in the health and educational domains---two areas that are of significance in the so-called developing world.

The following members are actively working in this area:

Lighton Phiri
Academic/Faculty

Robert M'sendo
Masters Student

Halwiindi Mulomba
Undergraduate Student

Astence Chinyama
Undergraduate Student

Memory Banda
Undergraduate Student

Kelly Mwitwa
Undergraduate Student

Hangoma Maambo
Undergraduate Student

Digital Libraries

The field of Digital Libraries (DLs) generally involves the study of digital collections of information and corresponding network-based services used to retrieve data from the collections. DLs are in effect information systems that are used to persistently store digital objects, manage the digital objects and, facilitate access to digital objects.

Our focus in the field of DLs, as a research group, mostly involves experimenting with techniques that can potentially facilitate efficient and effective access to digital objects stored in DLs.

The following members are actively working in this area:

Lighton Phiri
Academic/Faculty

Dokowe Tembo
Masters Student

Robert M'sendo
Masters Student

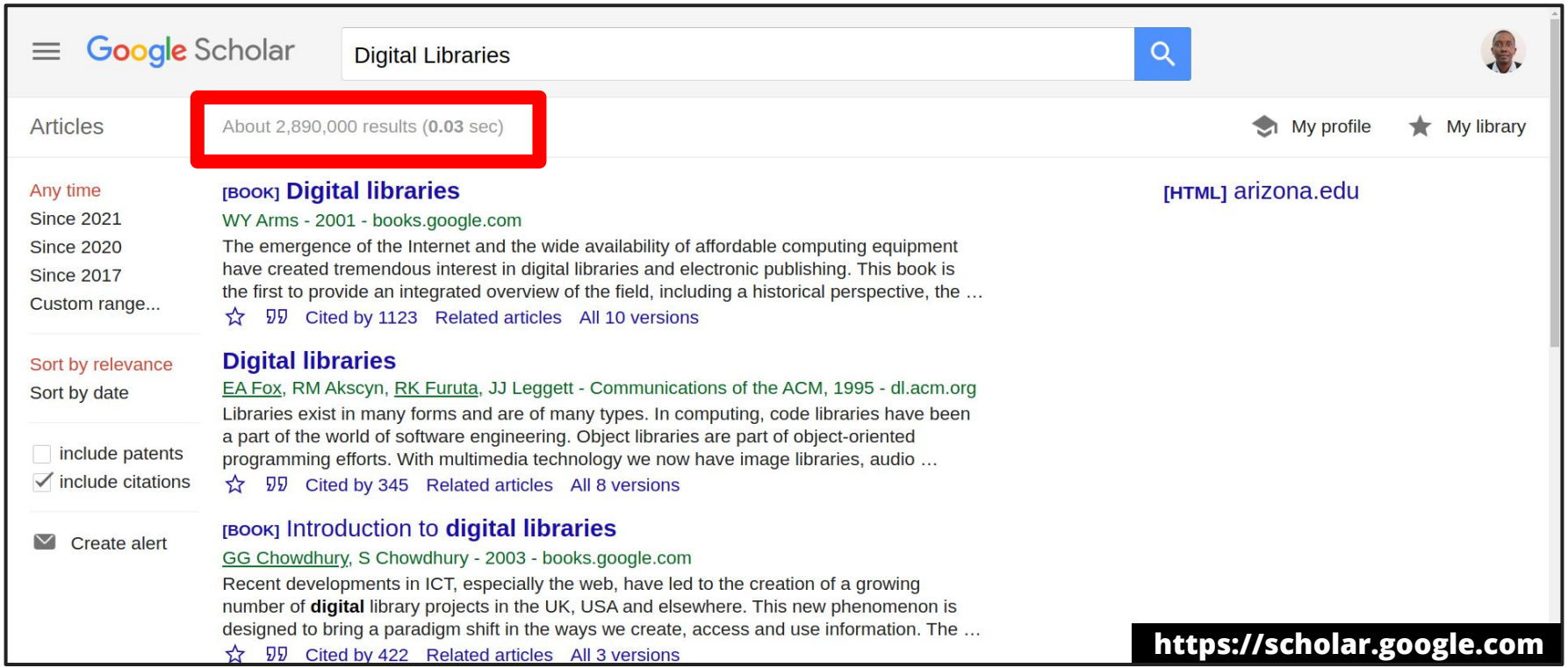
Angela Banda
Masters Student

http://datalab.unza.zm

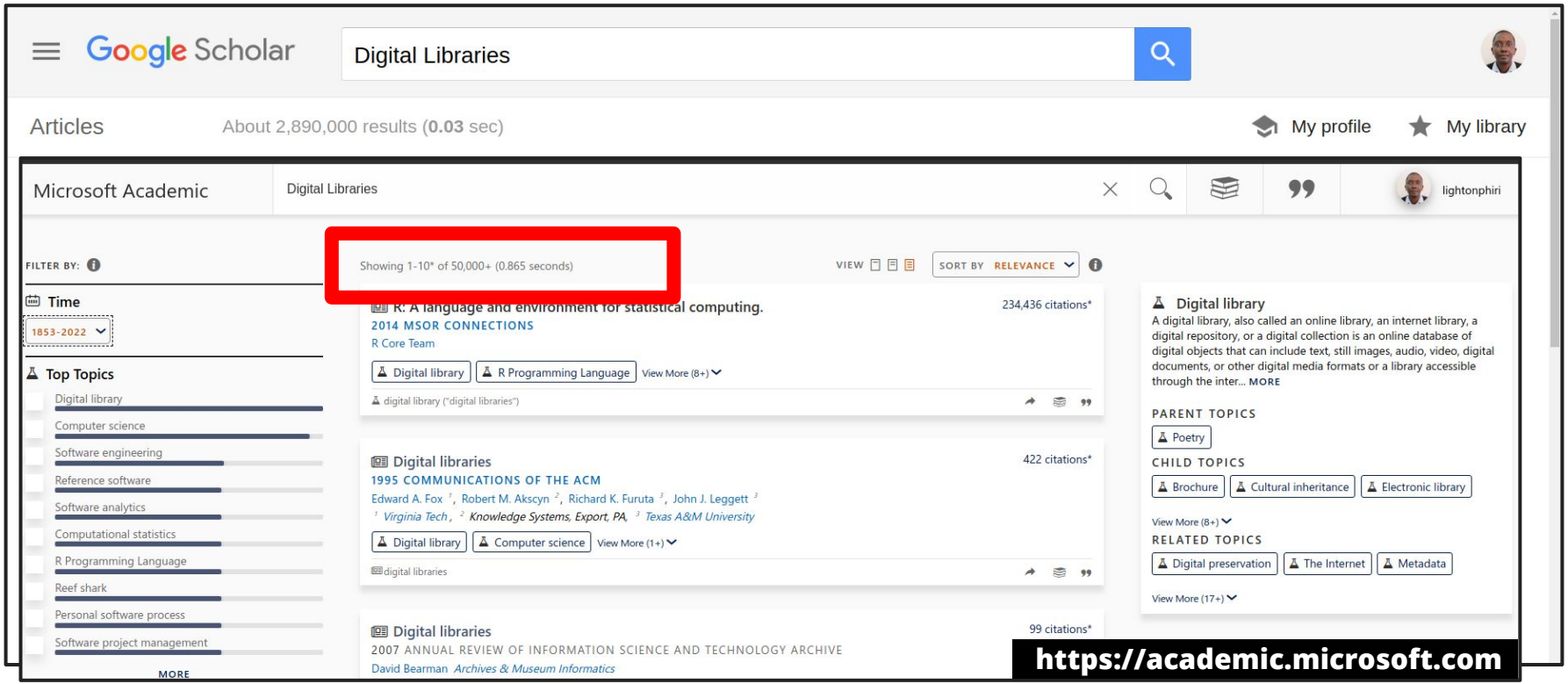
Outline

- Motivation
- Problem Statement
- Methodology
- Results and Discussion
- Conclusion and Future Work

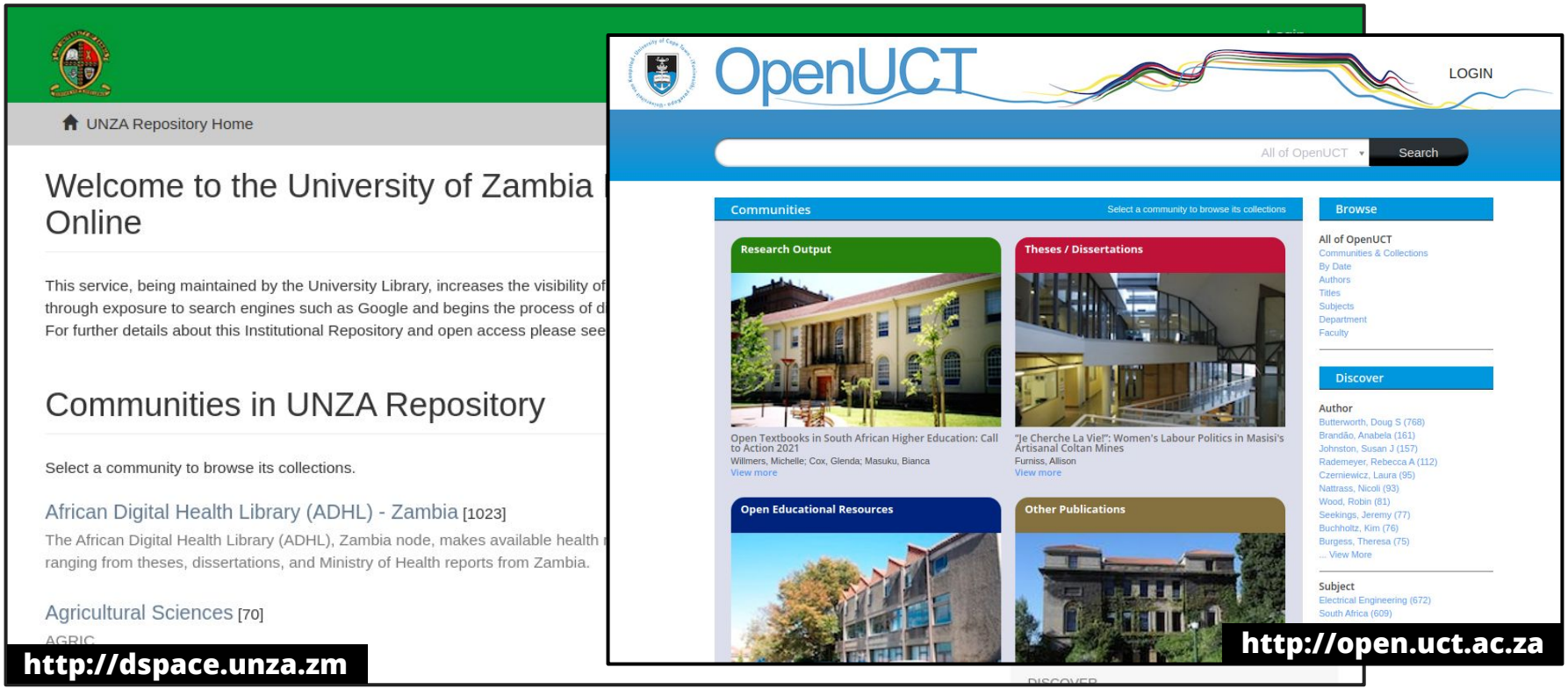
There is an Ever Increasing Amount of Scholarly Research Generated



There is an Ever Increasing Amount of Scholarly Research Generated

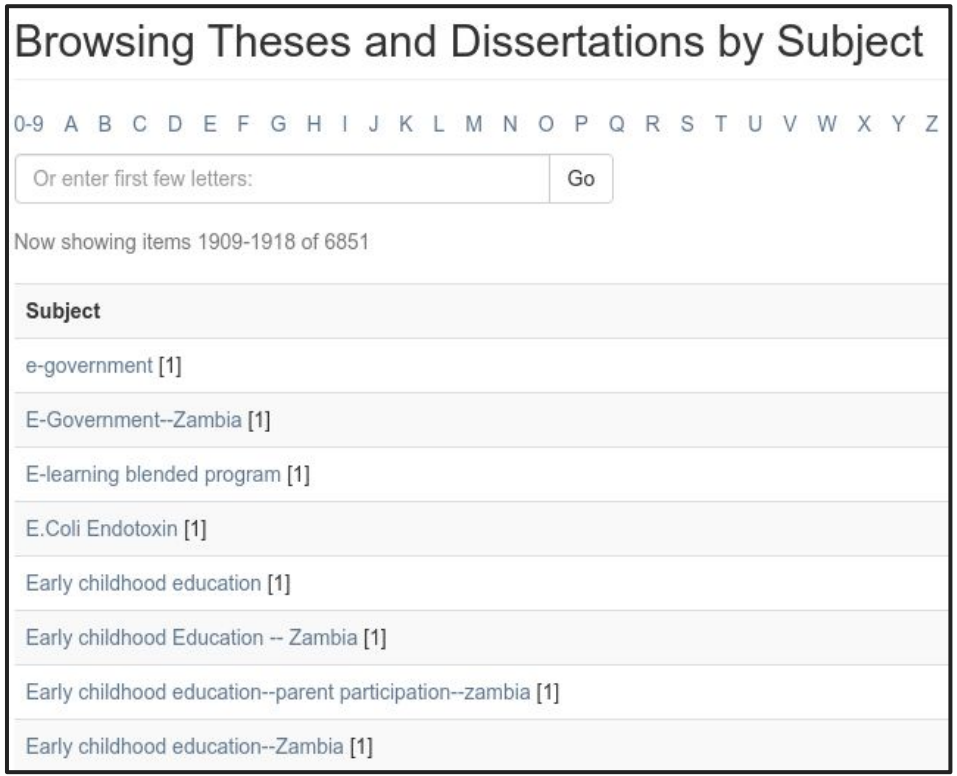


Discoverability Services Facilitate Findability of Scholarly Research in IRs



Problem Statement

- **There are numerous inconsistencies with digital object metadata elements used to describe subjects**
  - Lack of use of controlled vocabularies and subject headings compromises effective searching and browsing of scholarly research



Browsing Theses and Dissertations by Subject

0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

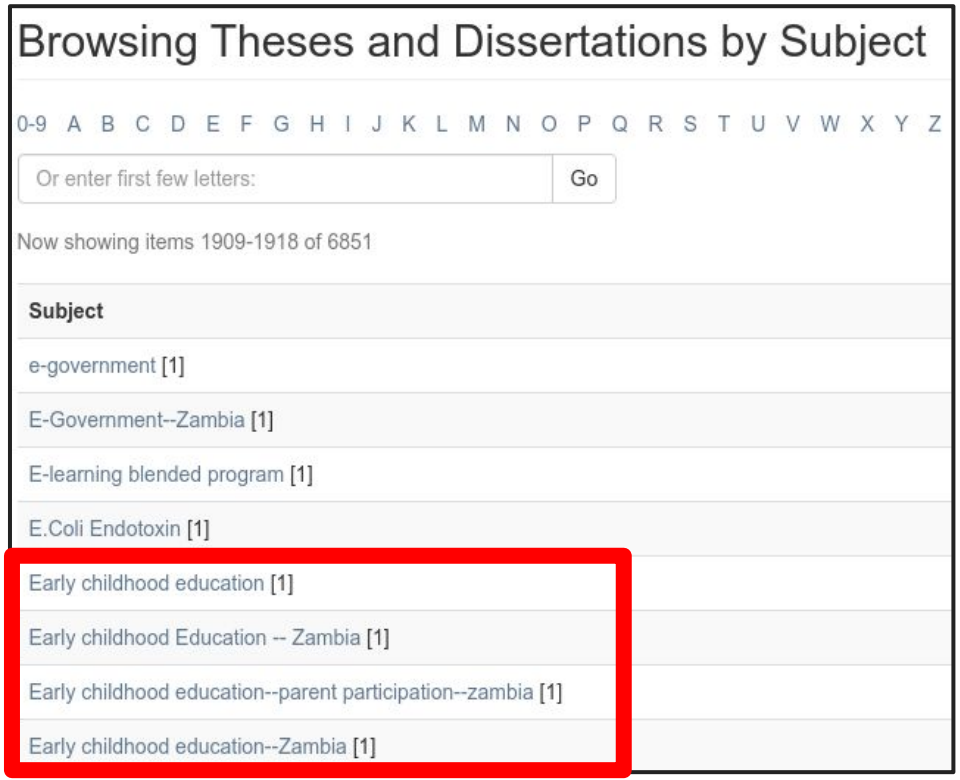
Or enter first few letters: Go

Now showing items 1909-1918 of 6851

Subject
e-government [1]
E-Government--Zambia [1]
E-learning blended program [1]
E.Coli Endotoxin [1]
Early childhood education [1]
Early childhood Education -- Zambia [1]
Early childhood education--parent participation--zambia [1]
Early childhood education--Zambia [1]

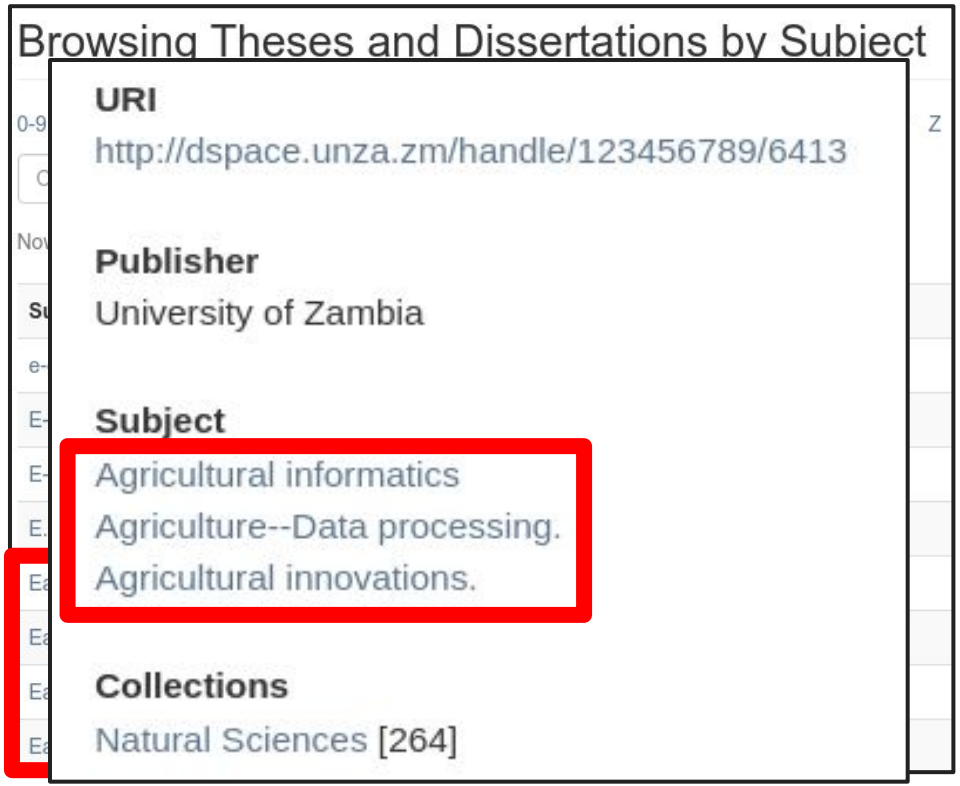
Problem Statement

- **There are numerous inconsistencies with digital object metadata elements used to describe subjects**
  - Lack of use of controlled vocabularies and subject headings compromises effective searching and browsing of scholarly research



Problem Statement

- **There are numerous inconsistencies with digital object metadata elements used to describe subjects**
  - Lack of use of controlled vocabularies and subject headings compromises effective searching and browsing of scholarly research



Methodology

- Situational analysis to determine the implications of non-use of controlled vocabularies
- Identification of subject-specific controlled vocabularies for various disciplines
- Usability study of IRs integrated with controlled vocabularies when compared with IRs without controlled vocabularies
- Implementation of multi-label subject classification model for classifying ACM CCS concepts and arXiv subjects

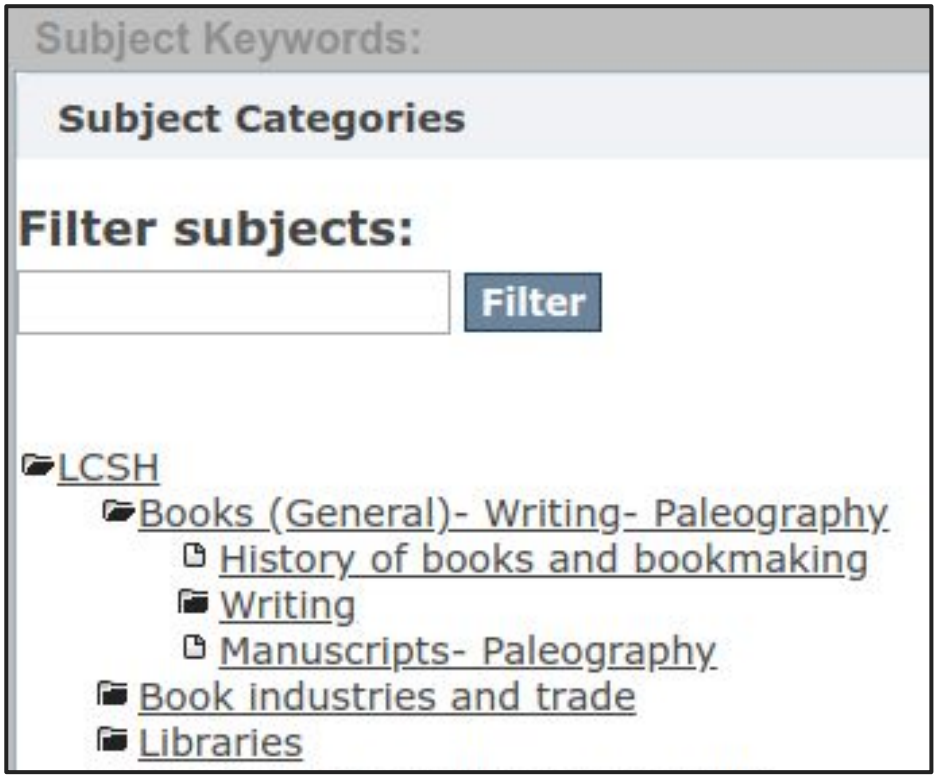
Methodology: Situational Analysis

- **Dublin Core encoded metadata harvested from three repositories**
  - NDLTD Union Catalog
  - University of Cape Town Computer Science Document Archive
  - University of Zambia Institutional Repository

```
<record>
 <header>
  <identifier>
   oai:dspace.unza.zm:123456789/6413
  </identifier>
  <datestamp>2020-09-21T10:38:06Z</datestamp>
  <setSpec>com_123456789_18</setSpec>
  <setSpec>col_123456789_84</setSpec>
 </header>
 <metadata>
  <oai_dc:dc>
   <dc:title>
    Automation of the grain purchasing Process for'
    Zambias food reserve Agency
   </dc:title>
   <dc:creator>Simukanga, Alinani</dc:creator>
   <dc:subject>Agricultural informatics</dc:subject>
   <dc:subject>Agriculture-Data processing.</dc:subject>
   <dc:subject>Agricultural innovations.</dc:subject>
   <dc:description>
    [...]
    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
```

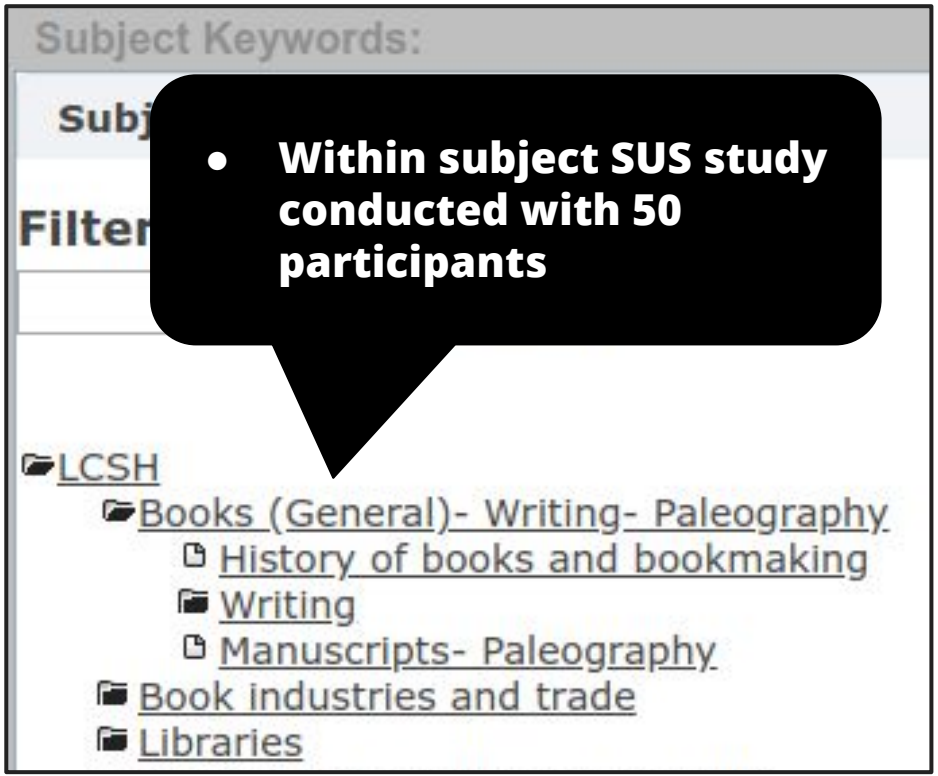
Methodology: Identification of Subjects and Usability Study

- 7 faculty interviewed to determine appropriate controlled vocabularies
- DSpace-powered IRs set-up to conduct controlled comparative study
  - IR #1: LCSH subjects
  - IR #2: No subjects
  - System Usability Scale used to assess usability



Methodology: Identification of Subjects and Usability Study

- 7 faculty interviewed to determine appropriate controlled vocabularies
- DSpace-powered IRs set-up to conduct controlled comparative study
  - IR #1: LCSH subjects
  - IR #2: No subjects
  - System Usability Scale used to assess usability



- Within subject SUS study conducted with 50 participants

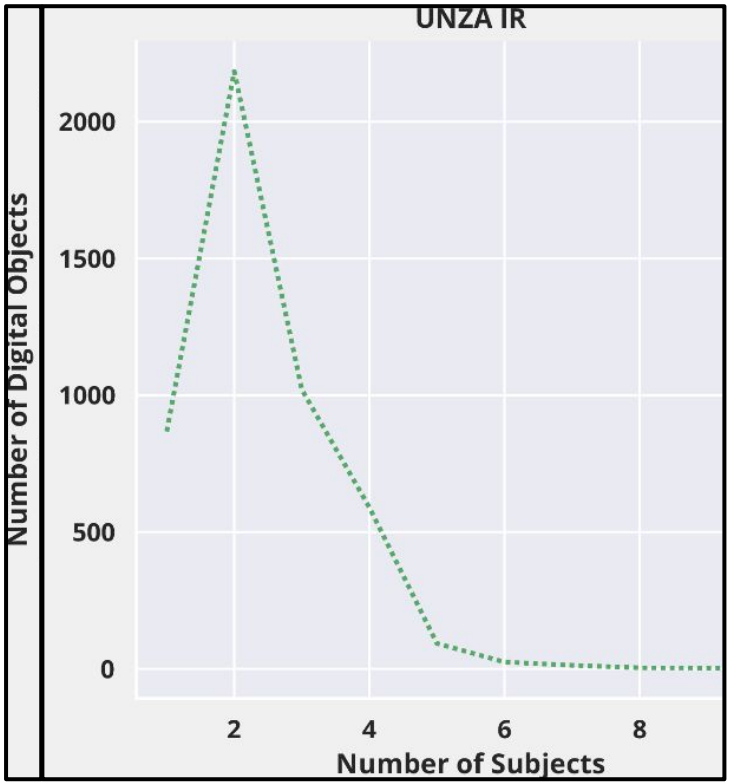
Methodology: Multi-label Subject Classifier

		Title + Abstract		
		F1 Score	Hamming Loss	Jaccard Score
SGDClassifier	TF-IDF	0.540	0.005	0.431
[...]	[...]	[...]	[...]	[...]

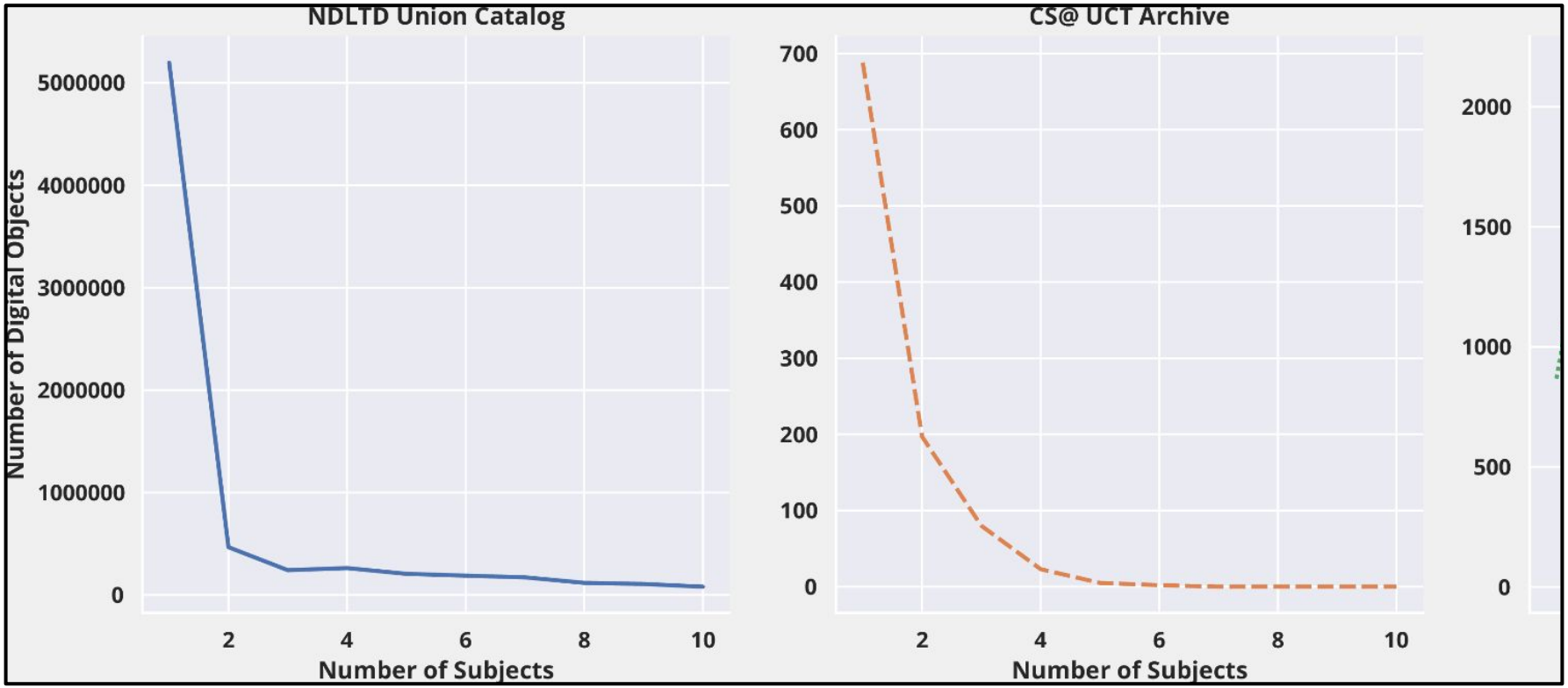
- Multi-label subject classification model implemented using arXiv CoRR dataset and validated using the UCT@ CS Document archive
  - Titles and Abstracts used as input features

Results and Discussion: Situational Analysis (2/2)

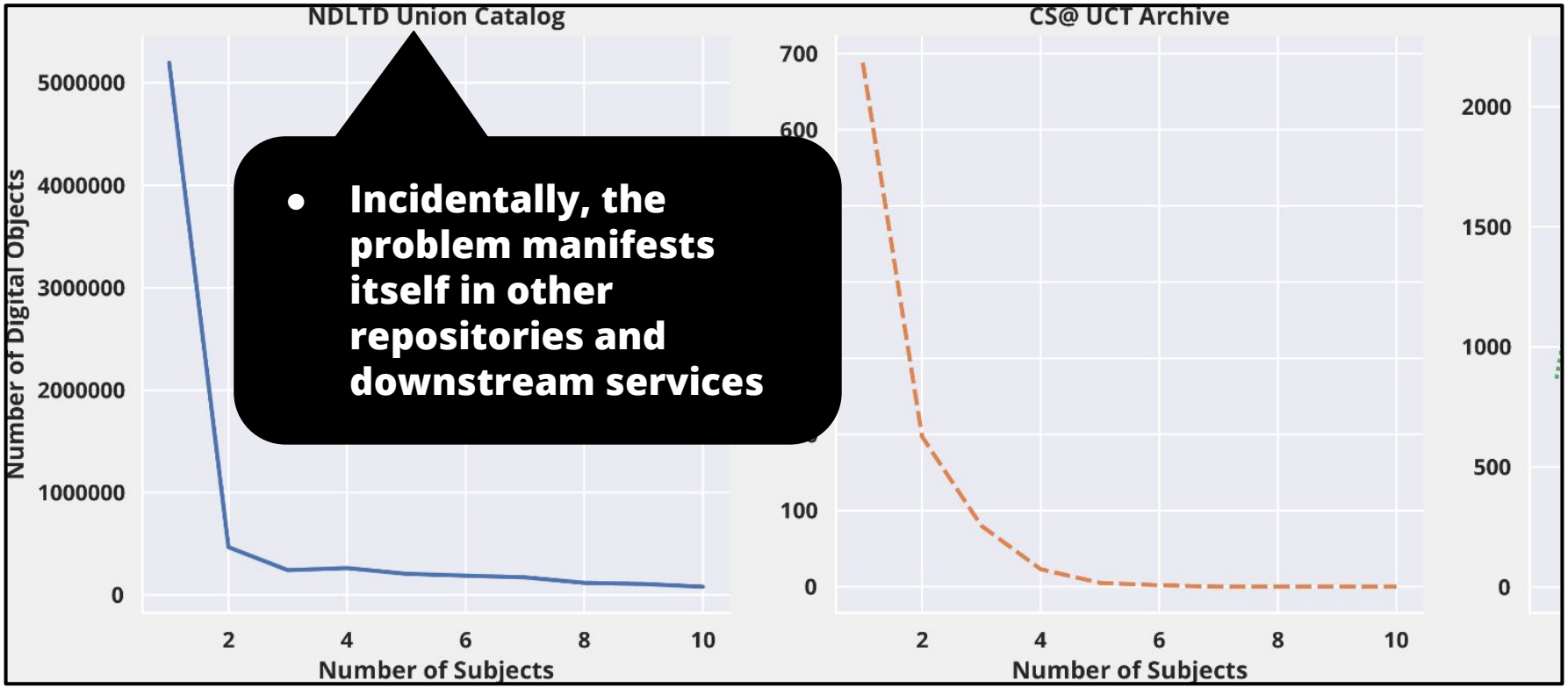
- Analysis 1. Metadata preparation and ingestion workflow based on internal policy
- Analysis 2. Subject heading usage is sparing. 92.1% of tags only associated with one publication
- Analysis 3. Domain-specific subject headings are not used. Internally devised LCSH used



Results and Discussion: Situational Analysis (1/2)

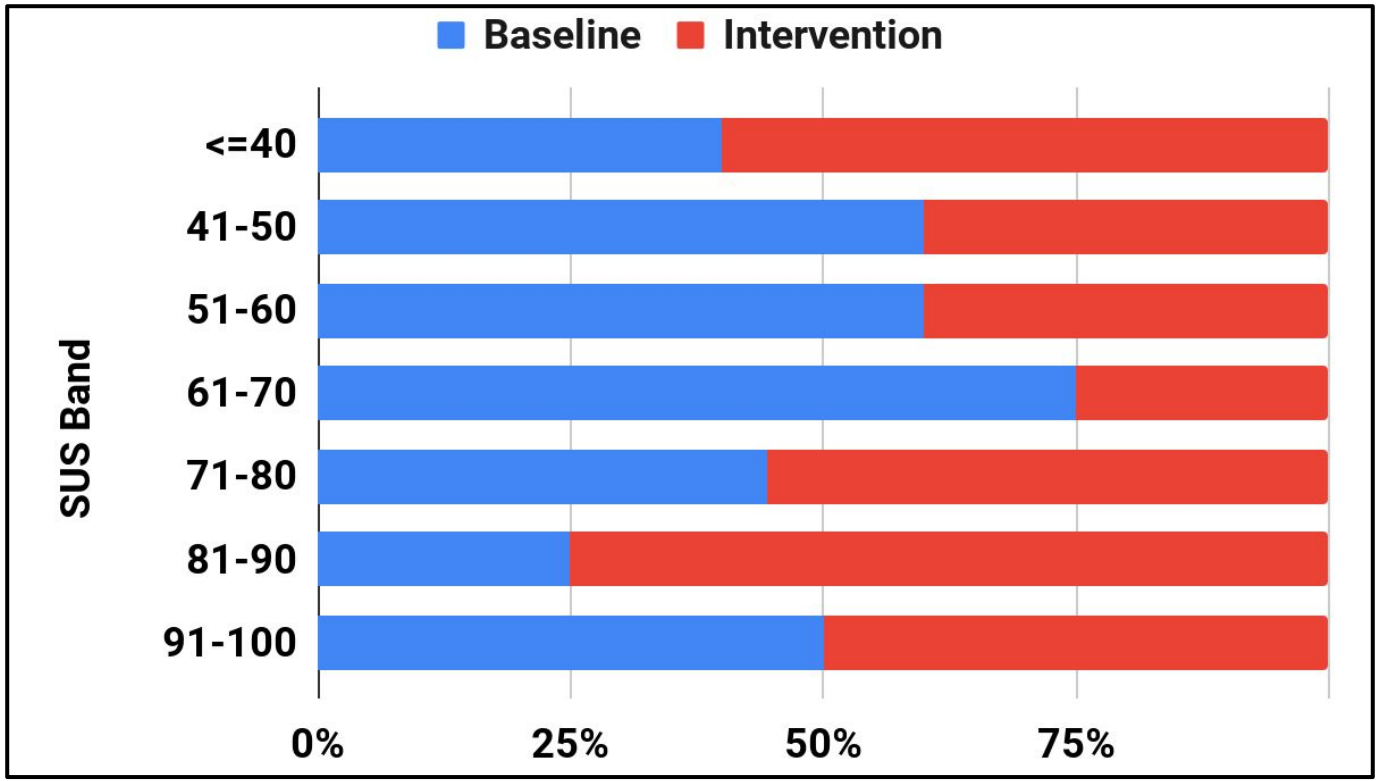


Results and Discussion: Situational Analysis (1/2)

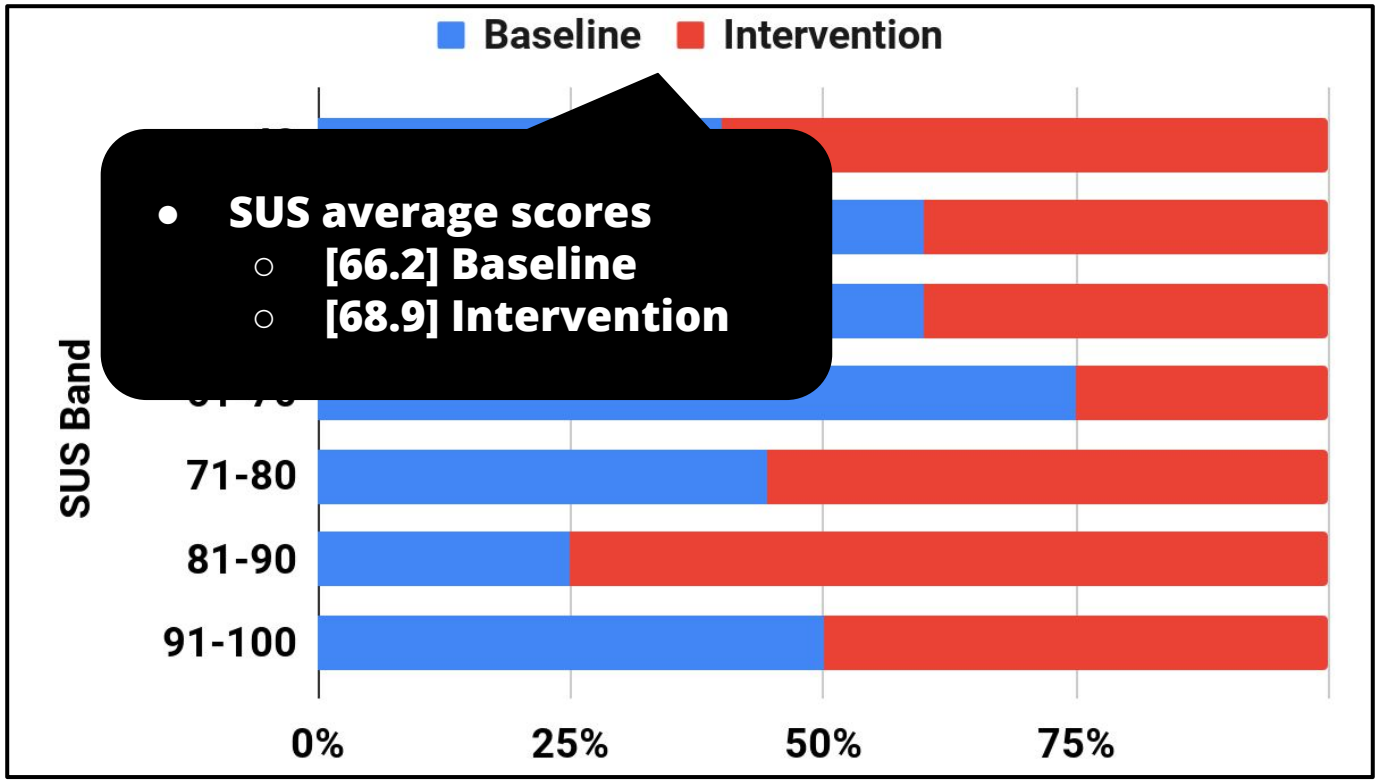


- **Incidentally, the problem manifests itself in other repositories and downstream services**

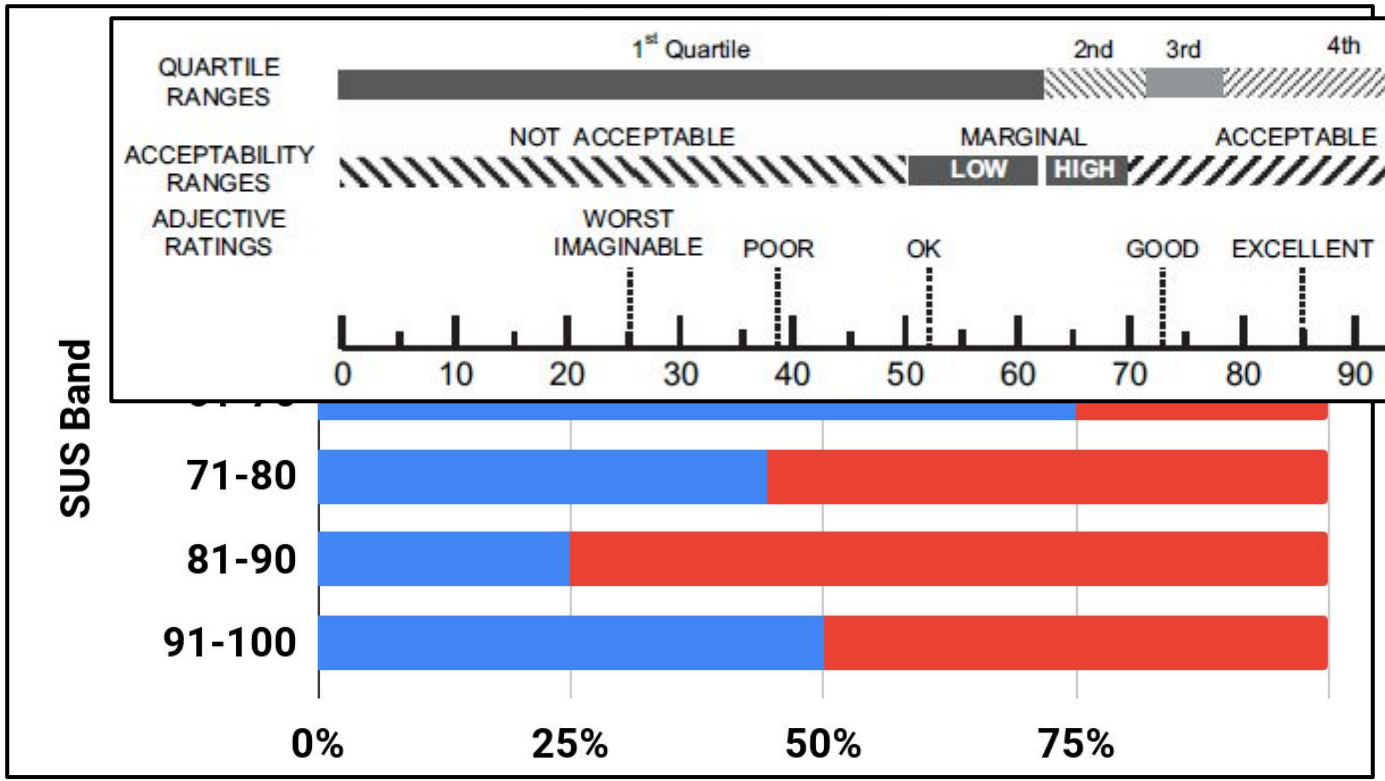
Results and Discussion: Comparative Analysis (1/2)



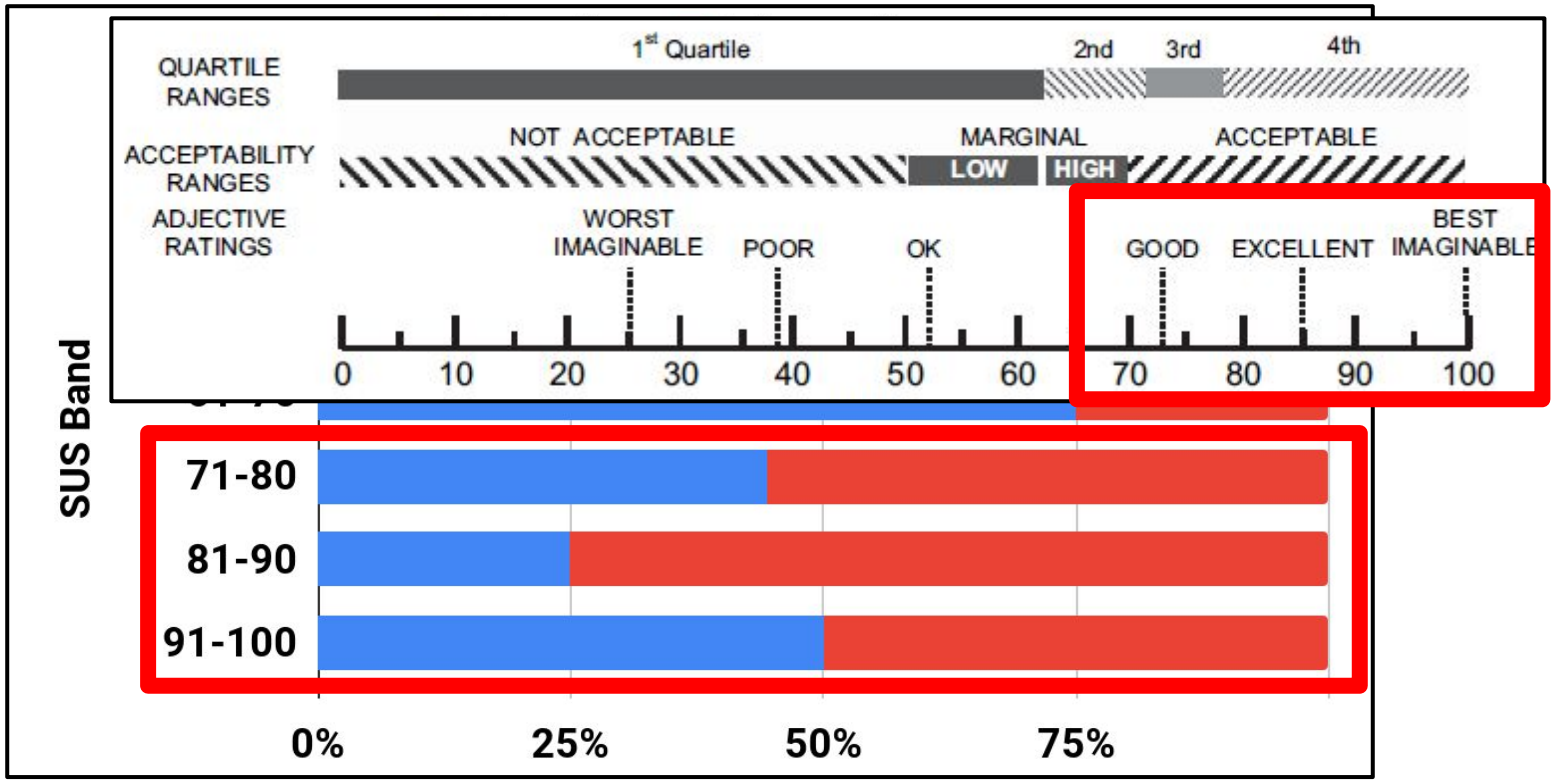
Results and Discussion: Comparative Analysis (1/2)



Results and Discussion: Comparative Analysis (2/2)



Results and Discussion: Comparative Analysis (2/2)

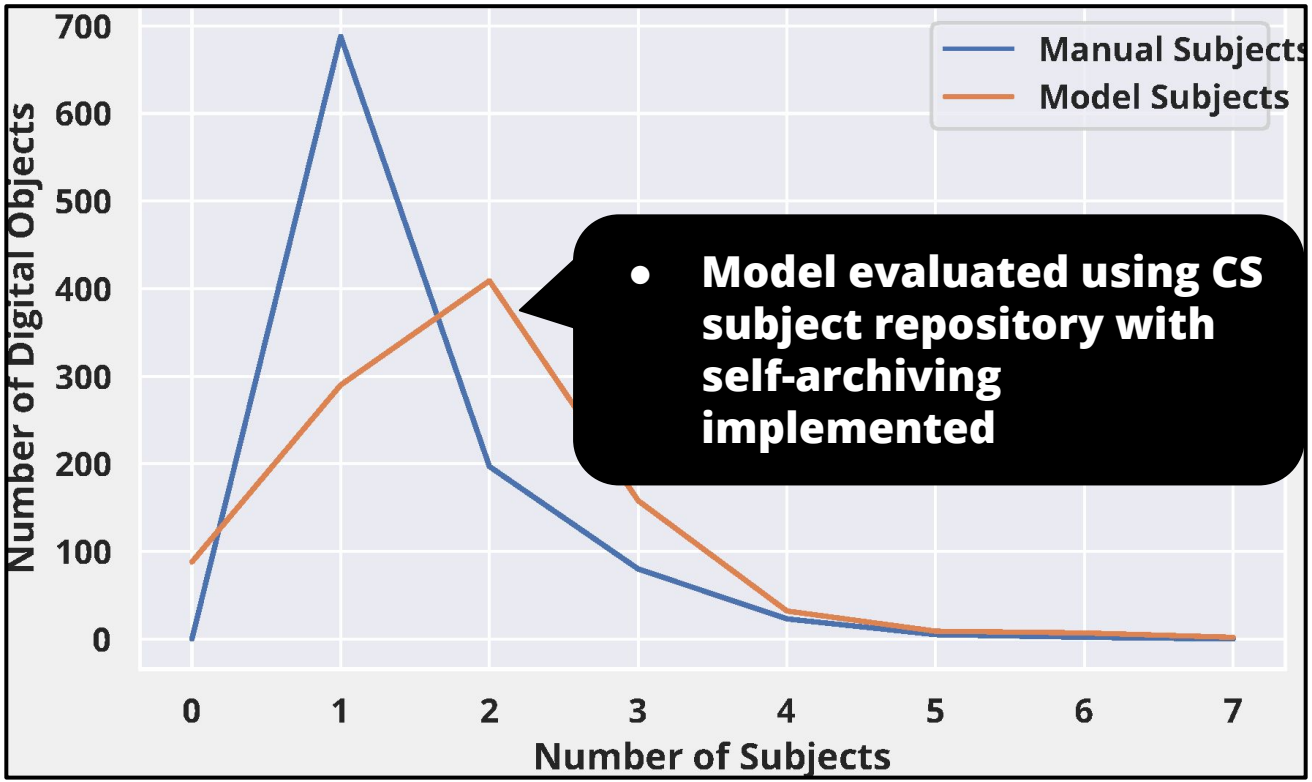


Results and Discussion: Multi-label Subject Classification Model—Implementation

		Title + Abstract		
		F1 Score	Hamming Loss	Jaccard Score
SGDClassifier	TF-IDF	0.540	0.005	0.431
[...]	[...]	[...]	[...]	[...]

- Approaches used: Binary Relevance, Classifier Chains and One-Versus-Rest
- Estimators: MultinomialNB vs SGDClassifier vs RandomForest
- Features: TF vs TF-IDF; Title vs Abstract vs Title + Abstract

Results and Discussion: Multi-label Subject Classification Model—Validation



Results and Discussion: Multi-label Subject Classification Model—Demonstration

C.2.4 · D.2.11 · F.1.1 · H.3.4 · H.3.5 · H.5.2

Computer Science - Artificial Intelligence · Computer Science - Computation and Language · Computer Science - General Literature · Computer Science - Human-Computer Interaction

- Six (6) arXiv and four (4) ACM CCS subject predicted by model

Conclusions and Future Work

- **Integrating IRs with subject controlled vocabularies can potentially complement self-archiving and, additionally, has the benefit ensuring that IRs are usable and effective**
- **Potential future work and/or direction**
  - Metadata cleaning, enhancement and augmentation of existing descriptive metadata
  - Implementation of subject classification models for other subject-specific controlled vocabularies
  - Automatic generation of subject classes for large-scale repositories such as the NDLTD Union Catalog

Q & A Session

- Comments, concerns and complaints?

Bibliography

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

[2] Phiri, L. (2020). A Multi-Faceted Multi-Stakeholder Approach for Increased Visibility of ETDs in Zambia. Cadernos BAD, (1). DOI: 10.1017/S0269888910000032

[3] 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, 14(3), 234-248. DOI: 10.1504/IJMSO.2020.112804

lighton.phiri@unza.zm
http://datalab.unza.zm
http://lis.unza.zm/~lightonphiri

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 · Lighton Phiri

University of Zambia
Lusaka, ZAMBIA

ACM/IEEE Joint Conference on Digital Libraries (JCDL 2021)
September 27–30, 2021