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

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

[BOOK] Digital libraries
WY Arms - 2001 - books.google.com
The emergence of the Internet and the wide availability of affordable computing equipment have created tremendous interest in digital libraries and electronic publishing. This book is the first to provide an integrated overview of the field, including a historical perspective, the ...☆ 99 Cited by 1123  Related articles  All 10 versions

Digital libraries
EA Fox, RM Akscyn, RK Furuta, JJ Leggett - Communications of the ACM, 1995 - dl.acm.org
Libraries exist in many forms and are of many types. In computing, code libraries have been a part of the world of software engineering. Object libraries are part of object-oriented programming efforts. With multimedia technology we now have image libraries, audio ...☆ 99 Cited by 345  Related articles  All 8 versions

[BOOK] Introduction to digital libraries
GG Chowdhury, S Chowdhury - 2003 - books.google.com
Recent developments in ICT, especially the web, have led to the creation of a growing number of digital library projects in the UK, USA and elsewhere. This new phenomenon is designed to bring a paradigm shift in the ways we create, access and use information. The ...☆ 99 Cited by 422  Related articles  All 3 versions

https://scholar.google.com
There is an Ever Increasing Amount of Scholarly Research Generated

https://academic.microsoft.com
Discoverability Services Facilitate Findability of Scholarly Research in IRs

Welcome to the University of Zambia Online

This service, being maintained by the University Library, increases the visibility of scholarship through exposure to search engines such as Google and begins the process of discoverability. For further details about this Institutional Repository and open access please see...

Communities in UNZA Repository

Select a community to browse its collections.

African Digital Health Library (ADHL) - Zambia [2023]
The African Digital Health Library (ADHL), Zambia node, makes available health information ranging from theses, dissertations, and Ministry of Health reports from Zambia.

Agricultural Sciences [70]

http://dspace.unza.zm

http://open.uct.ac.za
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
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
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
Methodology: Multi-label Subject Classifier

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature</th>
<th>F1 Score</th>
<th>Hamming Loss</th>
<th>Jaccard Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGDClassifier</td>
<td>TF-IDF</td>
<td>0.540</td>
<td>0.005</td>
<td>0.431</td>
</tr>
<tr>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
</tr>
</tbody>
</table>

- 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)

![Graph showing NDLTD Union Catalog and CS@UCT Archive](image)
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)

- SUS average scores
  - [66.2] Baseline
  - [68.9] Intervention
Results and Discussion: Comparative Analysis (2/2)

![Diagram showing SUS Band results](image)

- **71-80**:
  - 0% to 25%
  - 25% to 50%
  - 50% to 75%
- **81-90**:
  - 0% to 25%
  - 25% to 50%
  - 50% to 75%
- **91-100**:
  - 0% to 25%
  - 25% to 50%
  - 50% to 75%
Results and Discussion: Comparative Analysis (2/2)

[Diagram showing SUS Band with ranges 71-80, 81-90, 91-100]
Results and Discussion: Multi-label Subject Classification Model—Implementation

<table>
<thead>
<tr>
<th></th>
<th>Title + Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F1 Score</strong></td>
<td><strong>Hamming Loss</strong></td>
</tr>
<tr>
<td>SGDClassifier</td>
<td>0.540</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>[...]</td>
</tr>
<tr>
<td>[...]</td>
<td>[...]</td>
</tr>
</tbody>
</table>

- 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

- Model evaluated using CS subject repository with self-archiving implemented
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

URI
http://dspace.unza.zm/handle/123456789/6413

Publisher
University of Zambia

Subject
- Agricultural informatics
- Agriculture--Data processing.
- Agricultural innovations.

Collections
Natural Sciences [264]
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


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