Empirical Evaluation of ETD-ms Compliance for ETDs Harvested by the NDLTD Union Catalog

Adrian Chisale <adrian.chisale@unza.zm>
Cecilia Kasonde <19001415@student.unza.zm>
Lighton Phiri <lighton.phiri@unza.zm>

Department of Library and Information Science
University of Zambia
Lusaka, ZAMBIA

25th International Symposium on Electronic Theses and Dissertations (ETD 2022)
Novi Sad, Serbia | September 7–9, 2022
About Us

- 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

- Introduction
- Problem Statement
- Research Objectives
- Methodology
- Results and Discussion
- Conclusion and Future Work
Introduction: NDLTD Union Catalog (1/4)

- “The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of electronic theses and dissertations (ETDs)”.  
  - The global dissemination and preservation of ETDs is, in part, facilitated by the NDLTD Union Catalog
Introduction: NDLTD Union Catalog (2/4)

- Union Catalog harvests metadata from registered repositories
  - OAI-PMH protocol used to harvest data in Dublin Core format
- Union Catalog integrated with data provider

http://hdl.handle.net/10757/622568
Introduction: NDLTD Union Catalog (3/4)

- Union Catalog harvests metadata from registered repositories
  - OAI-PMH protocol used to harvest data in Dublin Core format
- Union Catalog integrated with data provider

http://hdl.handle.net/10757/622568
Introduction: NDLTD Union Catalog (3/4)

NDLTD Union Archive

About

This system collects metadata records for ETDs from institutions around the world and aggregates them into a single collection that can then be used by service providers. If you wish to search for ETDs, you can use the NDLTD Global ETD Search.

Recent Submissions

1. Surface Strain Measurement for Non-Intrusive Internal Pressure Evaluation of a Cannon
   Wed. 31 Aug 2022 05:59:08 UTC

<table>
<thead>
<tr>
<th>Institution</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Kentucky University Theses</td>
<td>3116</td>
</tr>
<tr>
<td>Wichita State University</td>
<td>1303</td>
</tr>
<tr>
<td>William and Mary</td>
<td>6245</td>
</tr>
<tr>
<td>Wirtschaftsuniversität Wien</td>
<td>3395</td>
</tr>
<tr>
<td>Worcester Polytechnic Institute</td>
<td>2034</td>
</tr>
<tr>
<td>Yale Medical student MD Thesis</td>
<td>122</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6291782</strong></td>
</tr>
</tbody>
</table>

http://union.ndltd.org/portal
Introduction: NDLTD Union Catalog (3/4)
Introduction: NDLTD Union Catalog (4/4)

- Union Catalog harvests metadata from registered repositories
  - OAI-PMH protocol used to harvest data in Dublin Core format
- Union Catalog integrated with data provider

http://hdl.handle.net/10757/622568
Introduction: NDLTD Union Catalog (4/4)

Search results
Showing 1 to 10 of 241 (0.14 seconds)
Spelling suggestions: "nhilton phiri" "nhilton chiri"

1. Simple Digital Libraries
Phiri, Lighton 01 August 2013 Ø (has links)
The design of Digital Library Systems (DLs) has evolved over time, both in sophistication and complexity, to complement the complex nature and sheer size of digital content being curated. However, there is also a growing demand from content curators, with relatively small-size collections, for simpler and more manageable tools and services to manage their content. The reasons for this particular need are driven by the assumption that simplicity and manageability might ultimately translate to lower costs of maintenance of such systems. This research proposes and advances for a minimalist and simplistic approach to the overall design of DLs. It is hypothesised that Digital Library (DL) tools and services based on such designs could potentially be easy to use and manage. A meta-analysis of existing DL and non-DL tools was conducted to aid the derivation of design principles for simple DLs. The design principles were then mapped to design decisions applied to the design of a prototype simple repository. In order to assess the effectiveness of the simple repository design, two

2. Investigating the Impact of Organised Technology-driven Orchestration on Teaching
Phiri, Lighton 01 October 2018 Ø (has links)
Orchestration of learning involves the real-time management of activities performed by educators in learning environments, with a particular focus on the effective use of technology. While different educational settings present unique problems, the common challenges have been noted to primarily be as a result of multiple heterogeneous activities and their associated intrinsic and extrinsic constraints. In addition to these challenges, this thesis argues that the complexities of orchestration are further amplified due to the ad hoc nature of the approaches and techniques used to orchestrate learning activities. The thesis proposes a streamlined approach to technology-driven orchestration of learning, in order to address these challenges and complexities. Specifically, the thesis proposes a framework that focuses on three core aspects of orchestration: activity management, resource management and sequencing.

http://search.ndltd.org
Problem Statement (1/3)

- While poor quality of ETD metadata records harvested has been cited as a longstanding issue, the full extent of the problem has not been explored
  - Suleman highlights the low adoption of the ETD-ms standard (Suleman, 2012).

Status: August 2012
As of 15 August 2012, there are 1982309 records in the Union Archive, from a total of 132 different sites and sub-collections.
The top contributors include OCLC (1228554), the IBICT Brazilian national collection (142252), Libraries and Archives Canada (121677) and the Australasian Digital Theses Program (56698).
All of these are themselves collections from multiple sites. The smaller contributors are individual institutions and, in a few rare cases, divisions within institutions.
Harvesting is triggered once a day at 3am UTC for all sites.

Analysis: Problems and Fixes
The following is a list of the problems encountered during the transition and how they were resolved.

- There is a need for greater automation in various aspects of the system in order to make it more scalable. For example, a repository administrator should be able to test and submit their site's baseURL or make changes to site information without direct human intervention. Errors should be reported to source repository administrators automatically.
- While ETDMS has been widely advertised, most sites still do not provide metadata in this format. The few sites that do provide ETDMS metadata use the older pre-2011 version of the standard instead of v1.1. The Union Archive could automatically translate ETDMS records and generate partial ETDMS records from DC records on-demand.
  - While the design philosophy of the Union Catalog project has always been to support external service providers, there is an inevitable delay before they update their indices. Some end-users have complained.

http://hdl.handle.net/10757/622568
Problem Statement (2/3)

Orchestration of learning involves the real-time management of activities performed by educators in learning environments, with a particular focus on the effective use of technology. While different educational settings present unique problems, the common challenges have been noted to primarily be as a result of multiple but complementary activities and their associated intrinsic and extrinsic constraints. In this context, it argues that the complexities of orchestration...
Problem Statement (2/3)

Developing an automated fall armyworm identification, early warning and monitoring system using a convolution neural network

http://dspace.unza.zm/handle/123456789/7141
Problem Statement (3/3)

ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations

version 1.1

http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html

Editors

Thom Hickey
Ana Pavani
Hussein Suleman

Outline

1. Introduction

### Problem Statement (3/3)

**ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations**

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>thesis.degree</td>
<td>Name of the degree associated with the work as it appears within the work. (example: Masters in Operations Research)</td>
<td>Optional, Repeatable</td>
</tr>
<tr>
<td>thesis.degree.name</td>
<td>Level of education associated with the document.</td>
<td>Optional, Repeatable</td>
</tr>
</tbody>
</table>

**Three levels are valid:**

- 0 Undergraduate (pre-masters)
- 1 Masters (pre-doctoral)
- 2 Doctoral (includes post-doctoral)

---

1. **Introduction**

Research Objectives

- Empirically evaluate metadata compliance to the ETD-ms metadata standard
  - NDLTD Union Catalog ETD metadata evaluation
  - Results could potentially inform how legacy metadata records could be rectified

- Understand the potential root causes for poor metadata harvested by the NDLTD Union Catalog
  - Results could potentially inform policy direction focused on improved metadata quality
Methodology (1/2)

- Metadata records harvested using the OAI-PMH protocol, using the oai_dc metadata prefix
  - No other existing format yielded desired results
Methodology (1/2)

Investigating the impact of organised technology-driven orchestration on teaching

Phiri, Lighton

Suleman, Hussein

Meinel, Christoph

Computer Science

Orchestration of learning involves the real-time management of activities performed by educators with particular focus on the effective use of technology. While different educational settings require various forms of technology, effective orchestration requires a comprehensive approach that integrates technology into the teaching process. The research aims to explore the effectiveness of technology-driven orchestration in enhancing teaching outcomes. The study will involve a mixed-methods approach, combining quantitative data from student performance metrics with qualitative insights from teacher feedback and student testimonials. This investigation will contribute to the understanding of how technology can be strategically employed to enhance the learning experience.
Methodology (2/2)

- UNZA Case Study
  - Empirical analysis of ETD metadata records
    - Distribution of Dublin Core elements
  - Analysis of IR policy
    - Document analysis of IR
  - Interviews with stakeholders
    - Four (4) IR policy makers
    - Three (3) IR content submitters

http://dspace.unza.zm
Results and Discussion: NDLTD (1/7)

- 7,563,684 metadata records harvested from 13,954 collections, using the oai_dc metadata prefix
- 4,901,643 metadata records used in analysis, after preprocessing
Results and Discussion: NDLTD (2/7)

<table>
<thead>
<tr>
<th>Metadata Element</th>
<th>Metadata Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc_contributor</td>
<td>64.27%</td>
</tr>
<tr>
<td>dc_coverage</td>
<td>96.16%</td>
</tr>
<tr>
<td>dc_creator</td>
<td>6.06%</td>
</tr>
<tr>
<td>dc_date</td>
<td>7.06%</td>
</tr>
<tr>
<td>dc_description</td>
<td>17.44%</td>
</tr>
<tr>
<td>dc_format</td>
<td>65.59%</td>
</tr>
<tr>
<td>dc_identifier</td>
<td>3.34%</td>
</tr>
<tr>
<td>dc_language</td>
<td>20.27%</td>
</tr>
<tr>
<td>dc_publisher</td>
<td>29.11%</td>
</tr>
<tr>
<td>dc_relation</td>
<td>0.39%</td>
</tr>
<tr>
<td>dc_rights</td>
<td>56.18%</td>
</tr>
<tr>
<td>dc_source</td>
<td>67.18%</td>
</tr>
<tr>
<td>dc_subject</td>
<td>25.22%</td>
</tr>
<tr>
<td>dc_title</td>
<td>3.03%</td>
</tr>
<tr>
<td>dc_type</td>
<td>27.36%</td>
</tr>
</tbody>
</table>
Results and Discussion: NDLTD (3/7)
Results and Discussion: NDLTD (4/7)

- 93% of non-null “dc:creator” values only have a single value
- Some multi-valued entries have publisher details, in addition to author details
  - Author details
  - Faculty details

<table>
<thead>
<tr>
<th>Values</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4299466</td>
<td>93</td>
</tr>
<tr>
<td>2</td>
<td>235837</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>29544</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>16480</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>8301</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4707</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2684</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2260</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1157</td>
<td>0</td>
</tr>
<tr>
<td>#</td>
<td>Print out multi-valued publishers</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td><code>print (var_creators) for var_creators in var_ndltd_dataset_inputCreators</code></td>
<td></td>
</tr>
</tbody>
</table>

Parks, Sophie Emma, University of Western Sydney, Faculty of Science and Technology
Tolar, Martin, University of Western Sydney, Macarthur Faculty of Business and Technology
Spriggs, Shelley, University of Western Sydney, Hawkesbury Faculty of Environmental Management and Agriculture
Hill, Geoff W., University of Western Sydney, Hawkesbury School of Social Ecology
Oka, Gusti Made, University of Western Sydney, Hawkesbury Faculty of Environmental Management and Agriculture
Falepau, David F., University of Western Sydney, Hawkesbury Faculty of Environmental Management and Agriculture
Boupha, Prasongsich C., University of Western Sydney, Hawkesbury Faculty of Science and Technology
Karikari, Steve, University of Western Sydney, Macarthur Faculty of Education
Toafoa, Tevita, University of Western Sydney, Hawkesbury Faculty of Agriculture and Horticulture
Weekley, Paul, University of Western Sydney, Hawkesbury Faculty of Environmental Management and Agriculture
Issouard, Godfrey, University of Western Sydney, Macarthur Faculty of Health
Yip, Hopi, University of Western Sydney, Hawkesbury Faculty of Science and Technology
Syafullah, University of Western Sydney, Hawkesbury Faculty of Science and Technology
Cathcart, Noel C., University of Western Sydney, Hawkesbury Faculty of Health, Humanities and Social Ecology
Simpson, Alyson Melanie, University of Western Sydney, Faculty of Humanities and Social Sciences
McDonald, Rodney, University of Western Sydney, Hawkesbury Faculty of Social Inquiry
Ross, Diianne S., University of Western Sydney, Hawkesbury Faculty of Science and Technology
Wallace, Gary E., University of Western Sydney, Hawkesbury Faculty of Environmental Management and Agriculture
Gleeson, Margaret McDonnell, University of Western Sydney, Hawkesbury Faculty of Health, Humanities and Social Ecology
Stewart, Brandon E., University of Western Sydney, Hawkesbury Faculty of Social Inquiry
Results and Discussion: NDLTD (5/7)

- 80% of non-null “dc:publisher” values only have a single value
- Random sample of records suggest entries primarily linked to “Institution”, “Faculty” and “Department”

<table>
<thead>
<tr>
<th>Values</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2784320</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>293082</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>101413</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>150250</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>127478</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1648</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>6165</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2559</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>4111</td>
<td>0</td>
</tr>
</tbody>
</table>
Results and Discussion: NDLTD (5/7)

1 #
2 #
3 # Print out multi-valued publishers
4 [print (var_publishers) for var_publishers in var_ndltd_dataset_input_publishers] [var_ndltd_dataset_input_publishers["dc_publisher"]]

1999 Adelaide Thesis (M.Mus.Perf.) -- University of Adelaide, Elder Conservatorium
Högskolan i Halmstad, Sektionen för Informationsvetenskap, Data- och Elektroteknik (IDE) Högskolan i Halmstad, Sektionen för Informationsvetenskap, Data- och Elektroteknik (IDE) Lethbridge, Alta.: University of Lethbridge, Dept. of Geography, c2013 Arts and Science Department of Geography
Lethbridge, Alta.: University of Lethbridge, Dept. of Geography, c2012 Arts and Science Department of Geography
Lethbridge, Alta.: University of Lethbridge, Dept. of Geography, c2013 Arts and Science Department of Geography
Lethbridge, Alta.: University of Lethbridge, Dept. of History, c2013 Arts and Science Department of History
KTH, Reglerteknik KTH, ACCESS Linnaeus Centre Stockholm
Högskolan i Halmstad, Sektionen för hälsa och samhälle (HOS) Högskolan i Halmstad, Sektionen för hälsa och samhälle (HOS) Linnéuniversitetet, Institutionen för organisation och entreprenörskap (OE) Linnéuniversitetet, Institutionen för organisation och entreprenörskap (OE) Linnaeus University, Institutionen för organisation och entreprenörskap (OE) Linnéuniversitetet, Institutionen för organisation och entreprenörskap (OE) Mälardalens högskola, Akademin för innovation, design och teknik Mälardalens högskola, Akademin för innovation, design och teknik Mälardalens högskola, Akademin för innovation, design och teknik Linnéuniversitetet, Institutionen för marknadsföring (MF) Linnéuniversitetet, Institutionen för marknadsföring (MF) Linnaeus University, Institutionen för marknadsföring (MF) Linnaeus University, Institutionen för marknadsföring (MF) Örebro universitet, Handelshögskolan vid Örebro Universitet Örebro universitet, Handelshögskolan vid Örebro Universitet Örebro universitet, Handelshögskolan vid Örebro Universitet Örebro universitet, Handelshögskolan vid Örebro Universitet Örebro universitet KTH, Reglerteknik KTH, ACCESS Linnaeus Centre Stockholm
800-1400°C. Où M représente la concentration molaire et T la température en K. Dans ces expériences sans volatiles nous avons criss Linköpings universitet, Optimeringslaboratorium Linköpings universitet, Tekniska högskolan Linköping
Linköpings universitet, Institutionen för beteendevetenskap och lärande Linköpings universitet, Filosofiska fakulteten Institutionen f Linköpings universitet, NISAL - Nationella institutet för forskning om äldre och äldrande Linköpings universitet, Filosofiska fakulteten Informa Healthcare Springer Nature
Results and Discussion: NDLTD (6/7)

- Less than 50% of “dc:contributor” values only have a single value
- While not always the case, the “dc:contributor” element primarily used to specify “Advisors/Supervisors”

<table>
<thead>
<tr>
<th>Values</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>738382</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>427397</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>267625</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>118617</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>118160</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>44555</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>16160</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>7412</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>7030</td>
<td>0</td>
</tr>
</tbody>
</table>
Results and Discussion: NDLTD (6/7)
## Results and Discussion: NDLTD (7/7)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.creator</td>
<td>93.00%</td>
<td>5.00%</td>
<td>1.00%</td>
<td>0.00%</td>
<td>1.00%</td>
</tr>
<tr>
<td>dc.contributor</td>
<td>42.00%</td>
<td>24.00%</td>
<td>15.00%</td>
<td>7.00%</td>
<td>11.00%</td>
</tr>
<tr>
<td>dc.date</td>
<td>70.00%</td>
<td>5.00%</td>
<td>13.00%</td>
<td>10.00%</td>
<td>3.00%</td>
</tr>
<tr>
<td>dc.description</td>
<td>45.00%</td>
<td>36.00%</td>
<td>6.00%</td>
<td>5.00%</td>
<td>8.00%</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>80.00%</td>
<td>8.00%</td>
<td>3.00%</td>
<td>4.00%</td>
<td>4.00%</td>
</tr>
<tr>
<td>dc.type</td>
<td>59.00%</td>
<td>29.00%</td>
<td>11.00%</td>
<td>1.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

- Most non-null multi-value elements are associated with a single value
Results and Discussion: Case Study (1/2)

![Bar Chart]

- **dc_contributor**: 100%
- **dc_coverage**: 100%
- **dc_creator**: 0.83%
- **dc_date**: 0.68%
- **dc_description**: 5.19%
- **dc_format**: 0.68%
- **dc_identifier**: 0.68%
- **dc_language**: 1.26%
- **dc_publisher**: 65.13%
- **dc_relation**: 100%
- **dc_rights**: 100%
- **dc_source**: 100%
- **dc_subject**: 1.2%
- **dc_title**: 0.68%
- **dc_type**: 1.17%

*September 7-9, 2022*  
*25th International Symposium on Electronic Theses and Dissertations (ETD 2022)*
## Results and Discussion: Case Study (2/2)

### 5.0 DIRECTIONS AND STRATEGIES

#### 5.1 Policy Direction
The IR shall contain all unpublished and published scholarly works and official documents of the University. The IR shall be used to store text, videos, images, audio and any other compatible digital format.

#### 5.2 Metadata Format for the Institutional Repository Materials
The items to be deposited in the IR shall have the following features:
- (a) Author
- (b) Title
- (c) Key words (subject headings)
- (d) Abstract
- (e) Full-text
- (f) Imprint (Publisher, place and year of publication)
- (g) Item type

### 6.0 INSTITUTIONAL FRAMEWORK

- **UNZA IR policy document analysis**
  - No metadata standard
  - Only five (5) Dublin Core elements specified

- **Interviews revealed that two (2) out of the four (4) Policy Makers interviewed were aware of ETD-ms.**
  - None of the ETD submitters were familiar with ETD-ms
Conclusions and Future Work

● Conclusions
  ○ Significant proportion of metadata records in NDLTD Union Catalog not compliant to ETD-ms standard
  ○ Policy direction and adoption of recommended metadata standards is crucial in ensuring that metadata is comprehensively

● Current and potential future work
  ○ Automatic generation of missing metadata for legacy ETD content
  ○ Detailed analysis of NDLTD Union Catalog ETD metadata records, focusing on repeatable elements
  ○ Development of IR policies and guidelines for improved metadata quality
Q & A Session

- Comments, concerns and complaints?
Bibliography


lighton.phiri@unza.zm
http://datalab.unza.zm
http://lis.unza.zm/~lightonphiri
Empirical Evaluation of ETD-ms Compliance for ETDs Harvested by the NDLTD Union Catalog

Adrian Chisale <adrian.chisale@unza.zm>
Cecilia Kasonde <19001415@student.unza.zm>
Lighton Phiri <lighton.phiri@unza.zm>

Department of Library and Information Science
University of Zambia
Lusaka, ZAMBIA

25th International Symposium on Electronic Theses and Dissertations (ETD 2022)
Novi Sad, Serbia | September 7-9, 2022