ETD Metadata / Union Catalog / IR Consortium Networks

USETDA 2018: 8th Annual Conference
Denver, Colorado; September 12, 2018

ETDs for Rookies

Edward A. Fox
Executive Director, Chairman of the Board
NDLTD, www.ndltd.org

fox@vt.edu  http://fox.cs.vt.edu/talks/2017
Professor, Department of Computer Science
Virginia Tech, Blacksburg, VA 24061 USA
Outline

• NDLTD: services, local work, connecting
• Digital Libraries: overview
• Metadata, OAI-PMH
• Background for Future Work
• Summary
Global ETD Search

Search the 5,005,212 electronic theses and dissertations contained in the NDLTD archive:

Type something to start searching...

**Advanced search tips**

The archive supports advanced filtering and boolean search.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject:</td>
<td>where the subject includes the word &quot;visualisation&quot;</td>
</tr>
<tr>
<td>title:</td>
<td>where the title includes the word &quot;computer&quot;</td>
</tr>
<tr>
<td>creator:</td>
<td>where the creator (author) is &quot;Hussein, Suleman&quot;</td>
</tr>
<tr>
<td>description:</td>
<td>where the description includes &quot;water rates&quot;</td>
</tr>
<tr>
<td>publisher:</td>
<td>where the publisher is &quot;McGill University&quot;</td>
</tr>
<tr>
<td>language:</td>
<td>where the language is &quot;english&quot;</td>
</tr>
<tr>
<td>apples AND</td>
<td>that contain both &quot;apples&quot; and &quot;bananas&quot;</td>
</tr>
<tr>
<td>bananas</td>
<td></td>
</tr>
<tr>
<td>apples NOT</td>
<td>that contain &quot;apples&quot; and do not contain &quot;bananas&quot;</td>
</tr>
</tbody>
</table>
Search results
Showing 1 to 10 of 1388 (0.071 seconds)
Spelling suggestions: "hussein suliman" "hussein sulene"

1. **MIRMaId: an interface for a content based Music Information Retrieval test-bed**
   Clotte, Candice Lynn 2006 (has links)
   Word processed copy. Includes bibliographical references (leaves 74-77). MIRMaId is an acronym for Music Information Retrieval Modular aid and is an interface that allows different content based retrieval tasks to be compared against each other to find optimal combinations of retrieval parameters for specialised problem domains. The dissertation describes the process of how the MIRMaId interface was developed, modified and refined.
   Information Technology

2. **Investigating the efficacy of XM and stylesheets to render electronic courseware for multiple learning styles**
   Du Toit, Masha 2007 (has links)
   Includes bibliographical references (leaves 87-90) The objective of this project was to test the efficacy of using Extensible Markup Language (XMI) - in particular the DocBook 5.0.05 schema - and Extensible Stylesheet Language Transformation (XSLT) to render electronic courseware that can be dynamically re-formatted according to a student's individual learning style. The text of a typical lesson was marked up in XML according to the DocBook schema, and several XSLT stylesheets were created to transform the XML document into different versions, each according to particular learning needs. These learning needs were drawn from the Felder-Silverman learning style model. The notes had links to trigger JavaScript functions that allowed the student to reformat the notes to produce different views of the lesson.
   Blood groups

3. **A digital library component assembly environment**
   Eyambe, Linda 2005 (has links)
   Includes bibliographical references (p. 105-113). Digital libraries (DLS), originally termed electronic libraries, are the result of the meshing of two communities: library professionals, including librarians and publishers, and computer scientists, including their poster child, Internet developers. Until recently, these two communities had very little interaction. Even now, it is not uncommon to find librarians who know little of digitisation or computer scientists who are unfamiliar with the tools of librarianship. The introduction of the Web has brought about greater collaboration between these two very different disciplines in a quest to facilitate the access to, and dissemination of, information.
   Computer Science

4. **Flexible packaging methodologies for rapid deployment of customisable component-based digital libraries**
   Mhlongo, Siyabonga 2006 (has links)
   Includes bibliographical references (leaves 82-86). Software engineering is a discipline concerned with manufacturing or developing software. Software plays a pivotal role in everyday life, an absence of which will be devastating to a number of governmental, recreational and financial activities, amongst many others. One of the latest branches of software engineering, component-based software engineering, is concerned with the development of software systems using already existing components which speculatively will ensure rapid and inexpensive software development processes. Parallel with the advances in software engineering, the field of digital libraries - a field dealing with Web-based access to and management of structured digital content - has adopted this development model from software engineering to shift focus from developing and using traditionally monolithic software systems to developing and using more flexible component-oriented software systems. Since componentised development approaches are relatively recent, other areas such as packaging and
   Computer Science

5. **Measuring the applicability of open data standards to a single distributed organisation: an application to the COMESA Secretariat**
Manage ETDs

This page provides resources for librarians, academic managers, and faculty to start and manage an electronic thesis program and to build and manage a thesis repository. We invite all NDLTD members and other interested parties to submit resources for inclusion in this web site. We also invite academic institutions to submit links to web pages and documents that may serve as valuable examples to others. Please send any contributions to admin@ndltd.org.

Start and Manage an ETD Program

- Set up an ETD Program
- Publishers and ETDs
- Guidance Documents for Lifecycle Management of ETDs

The Guidance Documents for Lifecycle Management of ETDs are a series of topical briefs to assist ETD stakeholders with the long-term curation and preservation of their ETDs. The cover the following issues:

- Introduction: Guidance Documents for the Lifecycle Management of ETDs
- Guidelines for Implementing ETD Programs - Roles and Responsibilities
- Guide to Access Levels and Embargoes of ETDs
- Briefing on Copyright and Fair Use Issues in ETDs
- Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs
- Managing the Lifecycle of ETDs: Curatorial Decisions and Practices
- Metadata for ETD Lifecycle Management
- Guide to ETD Program Planning and Cost Estimation
- Guide to Options for ETD Programs

The Lifecycle Management of Electronic Theses & Dissertations (ETDs) Workshop consists of a modular set of Creative Commons (CC-BY 4.0) licensed materials that are available for ETD programs or professional associations to make use of (and adapt) for educating a wide-range of ETD stakeholders. Modules include:

1. ETD Lifecycle Management Overview
2. ETD Lifecycle Guidance Documents Overview
3. ETD Lifecycle Management Tools Overview

The ETD Lifecycle Management Tools Manual provides resources and instruction to cover five major areas of lifecycle curation for ETDs, including:

1. Virus Checking
2. File Format Identification
3. Preservation Metadata
4. ETD SubmissionReference Link Archiving

- Examples of ETD Web Sites
- OhioLINK's Summary on Cataloging electronic resources, including ETDs
version 1.1

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

Editors

Thom Hickey
Ana Pavani
Hussein Suleman

Outline

1. Introduction
2. Authorities
3. Metadata Elements
   3.1 dc.title
   3.2 dc.creator
   3.3 dc.subject
   3.4 dc.description
   3.5 dc.publisher
   3.6 dc.contributor
   3.7 dc.date
   3.8 dc.type
   3.9 dc.format
   3.10 dc.identifier
   3.11 dc.language
   3.12 dc.coverage
   3.13 dc.rights
   3.14 thesis.degree
4. Global Qualifiers
5. Encodings and Crosswalks
Help Build Global ETD Search

This page is for ETD repository managers or those intending to start an ETD repository at their institution. Most students and researchers do not need to worry about this, except for prodding their libraries and graduate schools into action. Click here to use Global ETD Search.

NDLTD wants to build a comprehensive gateway to all doctoral and master-level theses and dissertations. This requires world-wide cooperation of all doctoral-degree granting institutions. As an organization dedicated to Open Access, NDLTD will harvest metadata from thesis repositories regardless of the participating universities' NDLTD membership status. Yet, NDLTD relies on membership dues to support this work. Please consider becoming an NDLTD member if access to thesis research is important to you.

To contribute content to Global ETD Search, a university, a research institution, or a consortium of universities must build a repository of their Electronic Theses and Dissertations (ETDs). For help in setting up a local repository, please consult Manage ETDs. It is the responsibility of local repository managers to develop acceptable open-access policies with their local communities. NDLTD encourages the maximum openness feasible: full text, all data, all metadata, and all multimedia files accessible to all with minimal embargo periods where needed.

NDLTD harvests metadata from participating repositories. It does not harvest text, data, figures, or multimedia. NDLTD organizes all harvested metadata in a NDLTD Union Archive. This Archive, which contains records of millions of ETDs, is managed by the University of Cape Town Digital Libraries Laboratory. The Union Archive itself is an Open Access resource, available to anyone wishing to build a service using the metadata stored in the Union Archive. Global ETD Search is a service built with data provided by the Union Archive.

To contribute metadata to the NDLTD Union Archive, you must configure the OAI-PMH interface of your ETD repository and you should register as an OAI data provider. Most repository software (DSpace, EPrints, Digital Commons, etc.) has a built-in software interface to allow harvesting of metadata by harvesters compliant with the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (OAI-PMH). If you must use a system (such as a custom-developed or in-house system) that does not have an OAI-PMH interface, the OAI website lists numerous tools that will help you to develop software to support the standard.

Finally, please e-mail the following details of your OAI registration to Hussein Suleman:

- name of your institution
- baseURL of the data provider
- list of set names that contain your ETDs, if your repository contains more than ETDs
- metadata formats that we should harvest (either oai_dc OR oai_dc+oai_etdms)

Click here for a list of institutions participating in the NDLTD Union Archive. A machine-readable version is available at http://union.ndltd.org/summary/.
Modern Dissemination Methods

- Social media
- WWW (and Web 2.0, 3.0)
- Digital libraries
  - Global tailored access, including through partners
  - Global access through search engines
  - Discovery, downloading, citing, re-usable science

- What aids discovery?
  - Title, abstract, categories
  - Figures (image search if separate)
  - Commenting, rating, tagging, recommending
Informal 5S & DL Definitions

DLs are complex systems that

• help satisfy info needs of users (societies)
• provide info services (scenarios)
• organize info in usable ways (structures)
• present info in usable ways (spaces)
• communicate info with users (streams)
DLs Support Many User Roles

Roles

Students → User
Readers → User
Teachers → User
Researchers → User
Advisors → User
Learners → User
Librarians → User

Digital Library
Why of Global Interest?

• **National projects** can preserve antiquities and heritage: cultural, historical, linguistic, scholarly

• Knowledge and information are essential to economic and technological **growth, education**

• DL - a **domain for international collaboration**
  – wherein all can **contribute** and **benefit**
  – which leverages investment in **networking**
  – which provides useful **content** on Internet & WWW
  – which will **tie nations and peoples together** more strongly and through **deeper understanding**
Digital Library Content

Content Types

- Text Documents
  - Articles, Reports, Books
- Video Audio
- Geographic Information
- Software, Programs
- Bio Information
- Images and Graphics
  - 2D, 3D, VR, CAT
- (Aerial) Photos
- Models Simulations
- Genome Human, animal, plant
<table>
<thead>
<tr>
<th>Infrastructure Services</th>
<th>Repository-Building</th>
<th>Add Value</th>
<th>Information Satisfaction Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creational</td>
<td>Conserving</td>
<td>Annotating</td>
<td>Browsing</td>
</tr>
<tr>
<td>Acquiring</td>
<td>Converting</td>
<td>Classifying</td>
<td>Collaborating</td>
</tr>
<tr>
<td>Cataloging</td>
<td>Copying/Replicating</td>
<td>Clustering</td>
<td>Customizing</td>
</tr>
<tr>
<td>Crawling (focused)</td>
<td>Emulating</td>
<td>Evaluating</td>
<td>Filtering</td>
</tr>
<tr>
<td>Describing</td>
<td>Renewing</td>
<td>Extracting</td>
<td>Providing access</td>
</tr>
<tr>
<td>Digitizing</td>
<td>Translating (format)</td>
<td>Indexing</td>
<td>Recommending</td>
</tr>
<tr>
<td>Federating</td>
<td></td>
<td>Measuring</td>
<td>Requesting</td>
</tr>
<tr>
<td>Harvesting</td>
<td></td>
<td>Publicizing</td>
<td>Surveying</td>
</tr>
<tr>
<td>Purchasing</td>
<td></td>
<td>Rating</td>
<td>Searching</td>
</tr>
<tr>
<td>Submitting</td>
<td></td>
<td>Reviewing (peer)</td>
<td>Visualizing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annotating</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Classifying</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clustering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluating</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extracting</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Indexing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Measuring</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Publicizing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rating</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reviewing (peer)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Surveying</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Translating</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(language)</td>
<td></td>
</tr>
</tbody>
</table>
Information Life Cycle (adapted)

Authoring
Modifying
Using
Classifying
Tagging
Recommending
Indexing
Storing
Retrieving
Distributing
Networking
Downloading
Discovering
Filtering
Distributing
Networking

Borgman et al. 1996
http://is.gseis.ucla.edu/research/dig_libraries/
Students and the Info. Life Cycle

- Learning
  - Respect for Teachers, Researchers
- Discovery
  - Excitement
- Documentation and reporting
  - Expression
- Dissemination
  - Fame and Fortune
- Community involvement
  - Service
- Community leadership
  - Compassion, Cooperation
Institutional Repositories

• “Institutional repositories are digital collections that capture and preserve the intellectual output of a single university or a multiple institution community of colleges and universities.”

• Crow, R. “Institutional repository checklist and resource guide”, SPARC, Washington, D.C., USA

• www.sparc.arl.org/resource/sparc-institutional-repository-checklist-resource-guide
For More Information (Examples)

- **Magazine**: www.dlib.org
  - Morgan Kaufmann: Witten... (several), Lesk (2nd edition)
  - M&C: https://sites.google.com/a/morganclaypool.com/dlibrary/ (4 books: foundations, key aspects, tech., appl.)
- **Conferences**
  - ICADL: http://www.icadl2017.org/
  - JCDL: http://www.jcdl.org
  - TPDL: http://www.tpdl2016.org/
- **Associations**
  - ASIS&T DL SIG: http://www.asis.org/SIG/dl.html
  - IEEE TCDL: http://www.ieee-tcdl.org
- **NSF**: http://www.dlib.org/dlib/july98/07griffin.html
Metadata Objects (MDOs)

- MARC
- Dublin Core -> ETD-MS
- RDF

- OAI (Open Archives Initiative): ORE, Protocol for Metadata Harvesting (PMH)

- Crosswalks, mappings
- Ontologies, Classification systems
Complex to Simple
(see DL Book 3 Ch. 1)

MARC ($50)

Dublin Core (DC) + thesis 19
OAI - Open Archives Initiative

• Advocacy for interoperability, from 1999
• Standard for transferring metadata among digital libraries
  – Protocol for Metadata Harvesting (PMH)
    • Simplicity
    • Generality
    • Extensibility
• Support for PMH => Open Archive (OA)
Open Archives Initiative

News from the OAI Community

• **mod_oai Project Aims at Optimizing Web Crawling.**
  The aim of the project is to create the mod_oai Apache software module that will expose content accessible from Apache Web servers via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). More information about the mod_oai project can be found at [http://www.modoai.org](http://www.modoai.org).

• **OLAC Archive on board European Space Agency mission.** On March 2, the Rosetta Disk left Earth on board an Ariane-5 rocket from the European Spaceport in Kourou, French Guyana. The materials on the disk come from the Rosetta 1000 Language Archive, an OLAC repository. (03/04).


---

Read Core Documents

• **Harvesting Protocol** (version 2.0)
  • 1.1 to 2.0 Migration Instructions
  • Implementation Guidelines
  • FAQ

Join the OAI Community

• Join OAI General list
  your email address

• Join OAI Implementers list
  your email address

Register as a data provider

Submit a post to our web site

Contact us:
openarchives@openarchives.org
OAI = Technical Umbrella for Practical Interoperability…

...that can be exploited by different communities
The World According to OAI

Service Providers

Discovery
Current Awareness
Preservation

Data Providers

Metadata harvesting
OAI – Repository Perspective

Required: Protocol

Set Structure

URI Scheme

MDO  MDO  MDO  MDO
MDO  MDO  MDO  MDO
DO   DO   DO   DO

Required: DC
OAI – Black Box Perspective
Early Union catalog: OCLC

- OCLC ran OAI data provider on TDs.
- Was getting data from WorldCat (so, from many sites!).
- Need DC and either ETD-MS or MARC.
- Has a set for ETDs.

- Now part of larger Union Catalog run at University of Cape Town (Hussein Suleman)
The **NDLTD Union Catalog** is an **OAI** server containing thesis and dissertation records harvested from registered NDLTD OAI repositories.

**NDLTD OAI baseURL**

http://alcme.oclc.org/ndltd/servlet/OAIHandler

**Protocol Request Forms**

1. Identify
2. GetRecord
3. ListIdentifiers *(Resumption)*
4. ListMetadataFormats
5. ListRecords *(Resumption)*
6. ListSets

**Notes**

- OAI XML responses from this repository include references to an XSL stylesheet. This allows modern web browsers (e.g. IE 6.0 or Netscape 7.0) to automatically render the XML into HTML instead of displaying the raw XML. Typical OAI harvester clients will naturally ignore the stylesheet reference and deal with the XML directly.
Metadata Specifications and Metadata Format: Completeness

- OCLC NDLTD Union catalog
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. Price and Quantity Effects of the German Real Estate Transfer Tax
   Fri, 04 Aug 2017 05:11:36 UTC
2. Wirtschaftswachstum und nachhaltiger Wohlstand. Mythen und Messung
   Fri, 04 Aug 2017 05:11:36 UTC
3. Satiated consumers: allocation of consumption time in an affluent society
   Fri, 04 Aug 2017 05:11:36 UTC
4. Nachhaltigkeit und Projektmanagement: Für eine breitere Perspektive auf Projekte
   Fri, 04 Aug 2017 05:11:36 UTC
5. The Illusion of Consumer Sovereignty in Economic and Neoliberal Thought
   Fri, 04 Aug 2017 05:11:36 UTC

Collection Statistics

<table>
<thead>
<tr>
<th>Collection</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBOLOGNA</td>
<td>0</td>
</tr>
<tr>
<td>UCF</td>
<td>6755</td>
</tr>
<tr>
<td>CRANFIELD</td>
<td>22</td>
</tr>
<tr>
<td>University of Oulu</td>
<td>5447</td>
</tr>
<tr>
<td>Arizona State University</td>
<td>6458</td>
</tr>
<tr>
<td>Atlanta University Center</td>
<td>3661</td>
</tr>
<tr>
<td>Australiasian Digital Theses Program</td>
<td>56698</td>
</tr>
<tr>
<td>Ball State University</td>
<td>3507</td>
</tr>
</tbody>
</table>
Background for Future Work

- Electronic Publishing
- Computer Literacy/Fluency -> Computational Thinking
- Scholarly Communication
- Open Access
- Institutional Repositories
- Open Archives Initiative (OAI)
Summary

• NDLTD: services, local work, connecting
• Digital Libraries: overview
• Metadata, OAI-PMH
• Background for Future Work
• Summary
Questions?
Discussion?
Recommendations?

Thank You!

fox@vt.edu
http://fox.cs.vt.edu