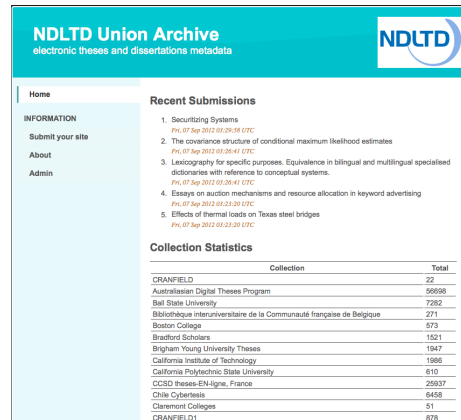


The NDLTD Union Catalog: Issues at a Global Scale



NDLTD Union Archive
electronic theses and dissertations metadata

Home

INFORMATION

- Submit your site
- About
- Admin

Recent Submissions

1. Securitizing Systems
Ph.D. Sep 2012 01:24:38 UTC
2. The covariance structure of conditional maximum likelihood estimates
Ph.D. Sep 2012 01:24:41 UTC
3. Lexicography for specific purposes. Equivalence in bilingual and multilingual specialised dictionaries with reference to conceptual systems.
Ph.D. Sep 2012 01:24:41 UTC
4. Essays on auction mechanisms and resource allocation in keyword advertising
Ph.D. Sep 2012 01:23:20 UTC
5. Effects of thermal loads on Texas steel bridges
Ph.D. Sep 2012 01:23:20 UTC

Collection Statistics

Collection	Total
CRANFIELD	22
Australasian Digital Theses Program	56598
Ball State University	7292
Bibliothèque interuniversitaire de la Communauté française de Belgique	271
Boston College	573
Bradford Scholars	1521
Bingham Young University Theses	1947
California Institute of Technology	11955
California Polytechnic State University	610
CCSD Theses (EN-ligne, France)	25937
Ohio Cyberthesis	8458
Claremont Colleges	51
CRANFIELD1	878



Hussein Suleman
hussein@cs.uct.ac.za

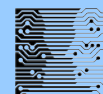
University of Cape Town
Department of Computer Science
Digital Libraries Laboratory

September 2012



Union Catalog Project

- Collect metadata for ETD collections internationally into a single collection, primarily using OAI-PMH
- Use the collection to
 - provide discovery services
 - showcase ETD production
 - encourage participation in ETD projects
- One Archive (at UCT), Many Catalogues (at VTLS, Scirus, etc.)





NDLTD Union Archive

NDLTD

NDLTD Union Archive

electronic theses and dissertations metadata

[Home](#)

INFORMATION

[Submit your site](#)

[About](#)

[Admin](#)

Recent Submissions

1. Securitizing Systems
Fri, 07 Sep 2012 03:29:58 UTC
2. The covariance structure of conditional maximum likelihood estimates
Fri, 07 Sep 2012 03:26:41 UTC
3. Lexicography for specific purposes. Equivalence in bilingual and multilingual specialised dictionaries with reference to conceptual systems.
Fri, 07 Sep 2012 03:26:41 UTC
4. Essays on auction mechanisms and resource allocation in keyword advertising
Fri, 07 Sep 2012 03:23:20 UTC
5. Effects of thermal loads on Texas steel bridges
Fri, 07 Sep 2012 03:23:20 UTC

Collection Statistics

Collection	Total
CRANFIELD	22
Australasian Digital Theses Program	56698
Ball State University	7282
Bibliothèque interuniversitaire de la Communauté française de Belgique	271
Boston College	573
Bradford Scholars	1521
Brigham Young University Theses	1947
California Institute of Technology	1986
California Polytechnic State University	610
CCSD theses-EN-ligne, France	25937
Chile Cybertesis	6458
Claremont Colleges	51
CRANFIELD1	878





VTLS Discovery Service

The screenshot displays the VTLS Discovery Service interface in a browser window. The page title is "Search | Driven by Chivas - (Private Browsing)". The URL is "thumper.vtls.com:6090/search/query?term_1=suleman&theme=NDLTD". The interface features a navigation bar with "Login", "Cart", "Heading Search", and "Clear Session" options. A search bar contains the term "suleman" and a "Search" button. Below the search bar, there are sections for "Account Login", "Refine your search", and "Current Search: suleman". The "Current Search" section shows "Results 1 to 17 of 17" and a "Sort by" dropdown set to "Relevance". A list of search results is displayed, each with a checkbox, a title, and the author's name. The results include:

- 1. **La production et valorisation des compétences sur le marché du travail : des approches néo-classiques à l'économie des conventions**
A producao e valorizacao das competencias no mercado de Trabalho : das aborgagens neo-classicas a economia das convenções
Suleman, Fatima
View Source Record
- 2. **Representations of gender in prime-time television a textual analysis of drama series of Pakistan television**
Suleman, Saleha
View Source Record
- 3. **Open digital libraries**
Suleman, Hussein
View Source Record
- 4. **Design and optimization of parallel haptic devices : Design methodology and experimental evaluation**
Khan, Suleman
View Source Record
- 5. **Epidemiology of malaria in Punjab, Pakistan : a case study in a rural community near Lahore**
Suleman, Mohammad
View Source Record
- 6. **Intestinal calcium transport in the chicken**
Bhatti, Mohammad Suleman
View Source Record
- 7. **Influence of Firm Structure on Profitability in the U.S. Pulp and Paper Industry (1960-1998)**
Suleman, Kanwar Muhammad





Scirus Discovery Service

The screenshot shows a web browser window titled "SCIRUS ETD Search — NDLTD - (Private Browsing)". The address bar shows "www.ndltd.org/serviceproviders/scirus-etd-search". The page header includes the NDLTD logo and the text "NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS". Navigation links include "Find ETDs", "Submit ETDs", and "Manage ETDs".

The main content area is titled "SCIRUS ETD Search" and features a search bar with the query "Suleman (\"interoperability\")". Below the search bar, there are radio buttons for "ETDs" (selected) and "All of the scientific web". A "Search" button is present, along with a "powered by SCIRUS" logo.

The search results section indicates "Searched for: All of the words Suleman (\"interoperability\")" and "Found: 42 results". The results are sorted by "relevance" and "date".

Three search results are visible:

- [Organização e representação da informação na biblioteca digital de teses e dissertações da Universidade do Estado de Santa ...](#)
Jaqueline Costa Alves.Jul 2009
...and content of your items to ensure **interoperability**. It s believed that the...the repositories, thus improving **interoperability** with other digital libraries. In fact...improve this procedure in favor of **interoperability** and retrieval of information...
- [Open digital libraries](#)
Suleman, Hussein.Jan 2002
...Libraries Hussein **Suleman** Dissertation submitted...digital library, **interoperability**, system architecture...2002 Hussein **Suleman** Open Digital Libraries Hussein **Suleman** (ABSTRACT) Digital...type of system **interoperability** is encouraged...
- [Streams, Structures, Spaces,Scenarios, and Societies \(SS\): A Formal Digital Library Framework and Its Applications](#)
Gonçalves, Marcos André.Dec 2004
...DLs) are complex information systems and therefore demand formal foundations lest development efforts diverge and **interoperability** suffers. In this dissertation, we propose the fundamen- tal abstractions of Streams, Structures, Spaces, Scenarios...

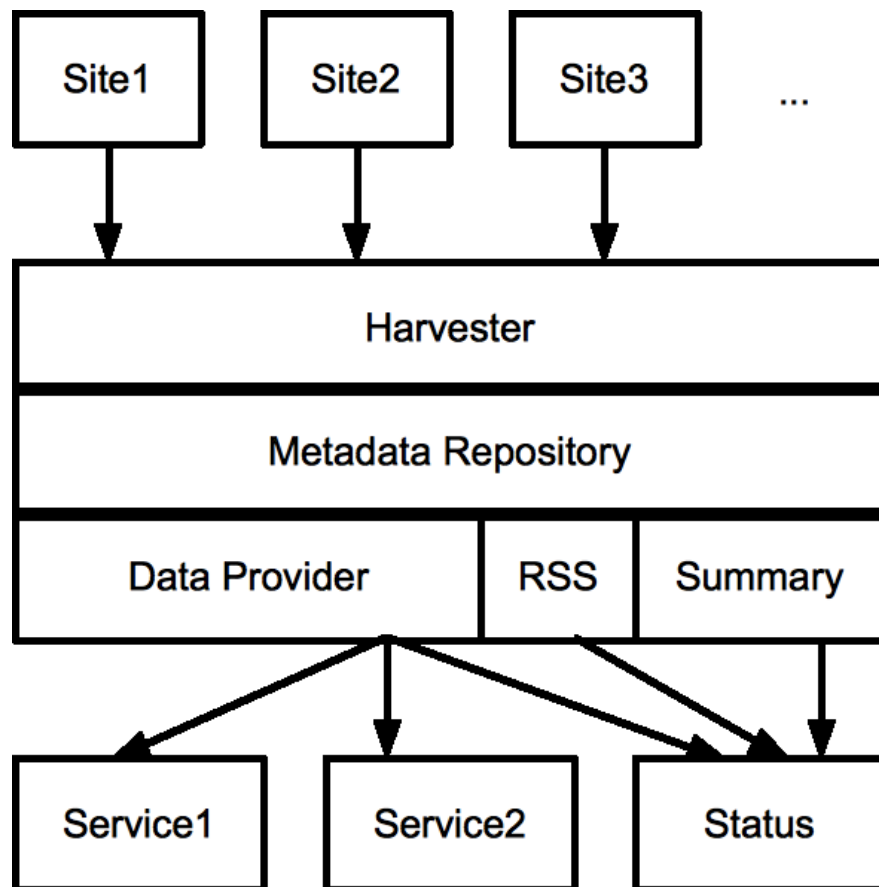
On the right side, there is a "Did you mean?" section with the suggestion [salesman interoperability](#). Below it, a "Refine using keywords found in the results:" section lists various keywords such as [customization](#), [digital library](#), [extensible](#), [greenstone](#), [heterogeneous data](#), [metadata](#), [object oriented query language](#), [relevance feedback](#), [schemas](#), [search engine](#), [semantic network](#), [service provider](#), [sistemas de](#), [spatial model](#), and [workflow](#).

The footer of the page includes "last modified 2008-09-15 09:26" and "Powered by Google Translate".





How Does It All Work?





OAI-PMH?

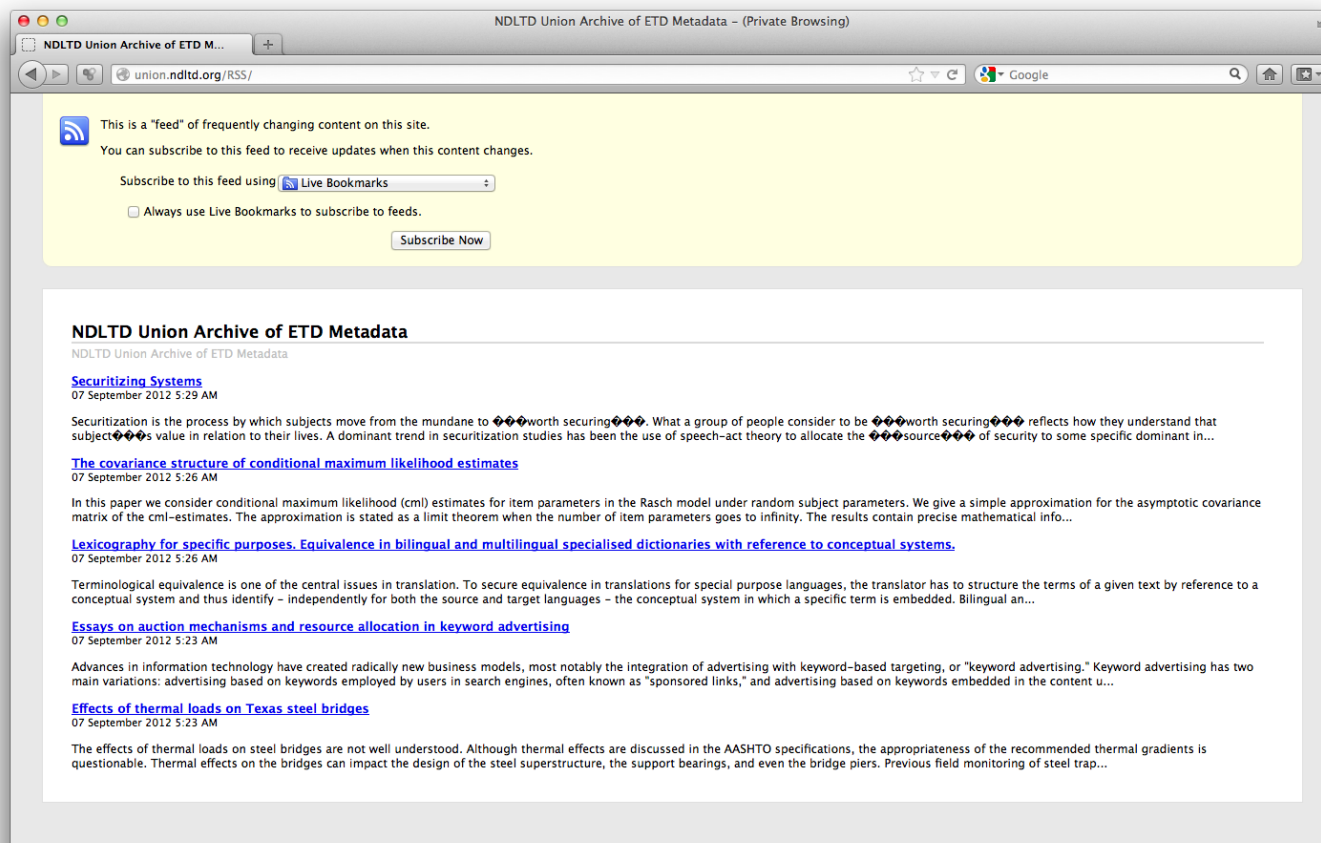
```
- <OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-09-07T07:21:33Z</responseDate>
  <request verb="Identify">http://union.ndltd.org:8080/union.OAI-PMH/</request>
  - <Identify>
    <repositoryName>NDLTD Union Archive of ETD Metadata</repositoryName>
    <baseURL>http://union.ndltd.org:8080/union.OAI-PMH/</baseURL>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>hussein@cs.uct.ac.za</adminEmail>
    <earliestDatestamp>2011-09-07T02:15:34Z</earliestDatestamp>
    <deletedRecord>persistent</deletedRecord>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
  - <description>
    - <eprints xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/eprints http://www.openarchives.org/OAI/1.1/eprints.xsd">
      - <content>
        <URL>http://union.ndltd.org/</URL>
        <text>NDLTD Union Archive of ETD Metadata</text>
      </content>
      <metadataPolicy/>
      <dataPolicy/>
    </eprints>
  </description>
</Identify>
</OAI-PMH>
```

<http://union.ndltd.org/OAI-PMH/>





RSS?



<http://union.ndltd.org/RSS/>





Quick Numbers

- 130 remote sites
- 2 metadata formats: DC, ETDMS
- 1986187 records
- largest sites:
 - OCLC >1million records from WorldCat
 - IBICT, LAC >100000 records each
- takes 22 hours to harvest from scratch!
- 1000 records per batch (instead of typical 100)





8 Problems and 6 Issues

- ❑ Global scale results in new problems
- ❑ Problems and issues are relevant to any large systems similar to Union Archive
- ❑ Problems can sometimes be resolved by better design at remote sites and often by redesign at central services

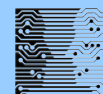




Problem 1: Globally Unique IDs

- Sites did not follow OAI identifier guidelines
 - that is, `oai:www.site.url:id_on_site`
- Instead some sites used:
 - `id_on_site`

- Solution:
 - all identifiers are encapsulated within:
 - `oai:union.ndltd.org:SITEID/id_on_site`

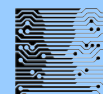




Problem 2: Globally Unique Site IDs

- Most sites are identified by an acronym
 - e.g., UB for University of Barcelona
- What about Botswana? Also UB!

- Solution:
 - all future site identifiers are FQDNs:
 - `www.ub.edu`
 - `www.ub.bw`

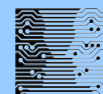




Problem 3: Orphan records

- Site with dead OAI data provider
- Records from defunct university
- Transitional records

- Approach:
 - Keep records unless explicitly removed





Problem 4: National site scale

- Large national sites have data providers that run off relational databases
- Insufficient indexing leads to slow responses from server
 - potentially decreasing in speed over time!
- Solution:
 - advise site managers on better database management
 - change timeouts per site on server





Problem 5: Union Archive scale

- 2 Millions records are quite a lot
- OAI-PMH best practices use batches of 100 records
 - too slow for large collections

- Solution:
 - Use larger batches – 1000 records for ~ 10 MB per response
 - Consider update to OAI-PMH!





Problem 6: Search Indices scale

- Portal software has a core indexer that is currently turned off
- Indexing fails because of poor organisation into batches

- Solution:
 - code has been rewritten
 - core indexing/searching will be activated in future





Problem 7: Multilingual support

- Software needs to support Unicode but this is still not the norm.

- Solution:
 - All tools were reconfigured to be in Unicode-mode
 - Mysql, Java servlets, etc.





Problem 8: Errors in records

- Still very much prevalent problem
- Records are not even valid XML!

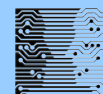
- Solution:
 - split responses into records before parsing
 - unparsable records are discarded
 - in future site admins could be contacted automatically





Issue 1: Deduplication

- Sites are included from:
 - OCLC subset
 - national/regional archives
 - directly
- Sometimes through 2-3 approaches
- De-duplication is still a known hard problem





Issue 2: Site maintenance

- ❑ Need tools to manage/curate collection
- ❑ Manual changes are difficult in a large collection
- ❑ Changes must propagate via OAI data provider





Issue 3: No Automation

- ❑ As collection grows, we need more automation
- ❑ Joining, error reports, etc. can be done mostly automatically
- ❑ Users can enter a baseURL and have it tested and queued without needing to contact a site admin





Issue 4: Metadata Standard

- All sites use DC
- Some sites use ETDMS v1.0
- None use ETDMS v1.1
- Need to move all sites to newer standard
 - Crosswalk old sites automatically
 - Start migration to new standard on old sites





Issue 5: Core Services

- Currently we rely exclusively on VTLS/Scirus
- A basic local search and browse service may be useful for immediate use and checking





Issue 6: Who is responsible?

- When records are invalid, who must fix them?
 - origin site?
 - regional site?
 - NDLTD?
 - all?
- Need better mechanisms for quality control





Conclusions

- In spite of all these issues, Union Catalog Project is working.
 - harvesting every day

- Still many unresolved issues.
- Most problems relate to scale and are relevant to other efforts as well.
- We still have much to learn about operating at larger scales...



that's all folks!



questions?

*search for "hussein suleman"
on Google or Facebook*

*and if you are the tweeting
#etd2012*