The NDLTD Union Catalog: Issues at a Global Scale

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

Recent Submissions

1. Securizing Systems
   Fri, 07 Sep 2012 01:29:58 UTC
2. The covariance structure of conditional maximum likelihood estimates
   Fri, 07 Sep 2012 01:29:41 UTC
3. Lexicography for specific purposes. Equivalence in bilingual and multilingual specialised dictionaries with reference to conceptual systems.
   Fri, 07 Sep 2012 01:29:41 UTC
4. Essays on auction mechanisms and resource allocation in keyword advertising
   Fri, 07 Sep 2012 01:30:06 UTC
5. Effects of thermal loads on Texas steel bridges
   Fri, 07 Sep 2012 01:29:20 UTC

Collection Statistics

<table>
<thead>
<tr>
<th>Collection</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRANFIELD</td>
<td>22</td>
</tr>
<tr>
<td>Australasian Digital Theses Program</td>
<td>50696</td>
</tr>
<tr>
<td>Ball State University</td>
<td>7382</td>
</tr>
<tr>
<td>Bibliothèque interuniversitaire de la Communauté française de Belgique</td>
<td>271</td>
</tr>
<tr>
<td>Boston College</td>
<td>573</td>
</tr>
<tr>
<td>Bradford Scholars</td>
<td>1521</td>
</tr>
<tr>
<td>Brigham Young University Theses</td>
<td>1947</td>
</tr>
<tr>
<td>California Institute of Technology</td>
<td>1986</td>
</tr>
<tr>
<td>California Polytechnic State University</td>
<td>610</td>
</tr>
<tr>
<td>UCSD Thesis-ENIGN, France</td>
<td>25937</td>
</tr>
<tr>
<td>Chile Cyberesis</td>
<td>6458</td>
</tr>
<tr>
<td>Claremont Colleges</td>
<td>51</td>
</tr>
<tr>
<td>CRANFIFI D1</td>
<td>47A</td>
</tr>
</tbody>
</table>
VTLS Discovery Service
Scirus Discovery Service

Search
Suleman ("interoporability")

Search for: All of the words Suleman ("interoporability")
Sort by: relevance | date

Organização e representação de 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...

Goncalves, Marcos Andre December 2004

...DLS) are complex information systems and therefore demand formal foundations lest development efforts diverge and Interoperability suffers. In this dissertation, we propose the fundamental abstractions of Streams, Structures, Spaces, Scenarios...

Developing semantic digital libraries using data mining techniques
Kim, Hyunki Jan 2005

Last modified 2008-03-25 19:29:20

Did you mean?
Suleman interoperability

Refine using keywords found in the results:
customization digital library extensible open-source heterogeneous data metadata object oriented query language relevance feedback schemas search engine semantic network service provider systems de spatial model workflow
How Does It All Work?
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 >1 million 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 ~10MB 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