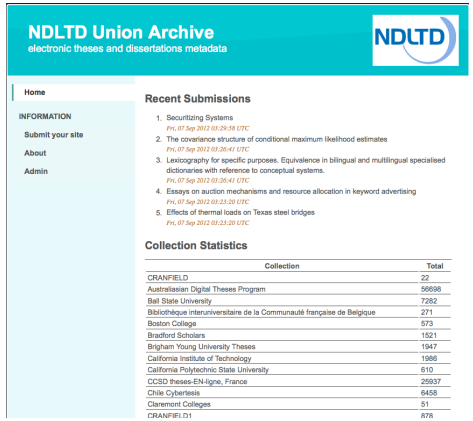
The NDLTD Union Catalog: Issues at a Global Scale



Hussein Suleman
hussein@cs.uct.ac.za

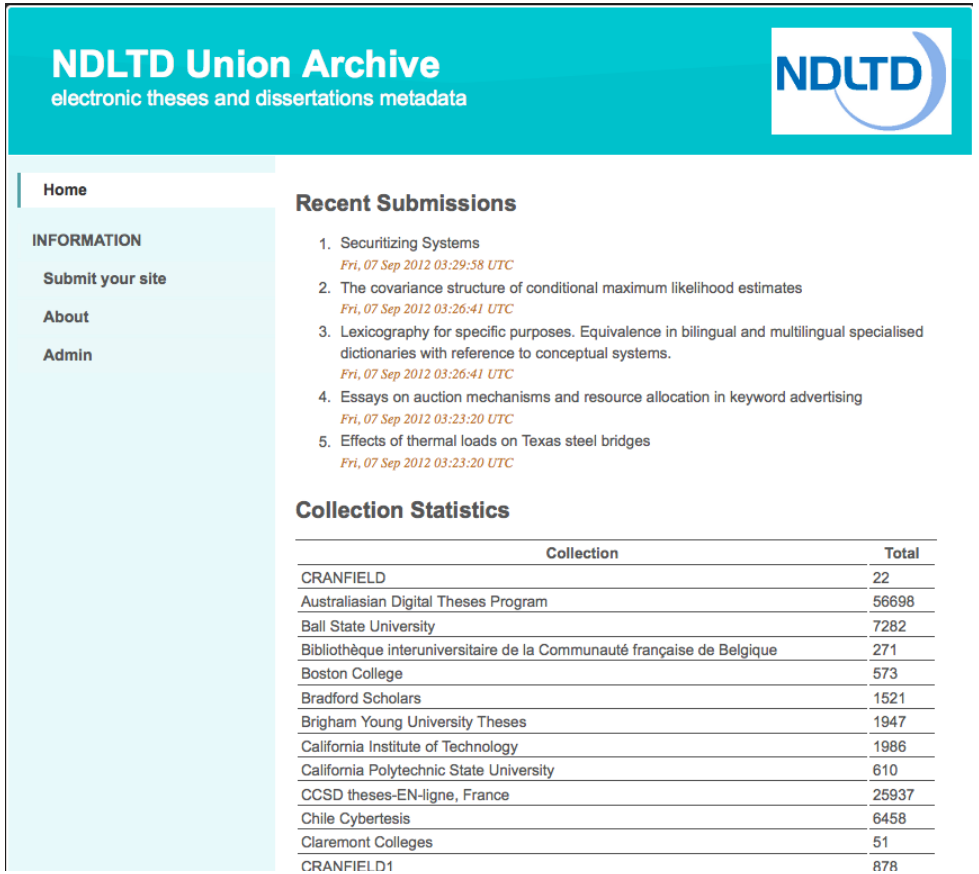
University of Cape Town
Department of Computer Science
Digital Libraries Laboratory

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



NDLTD Union Archive
electronic theses and dissertations metadata

NDLTD

Home

INFORMATION

Submit your site

About

Admin

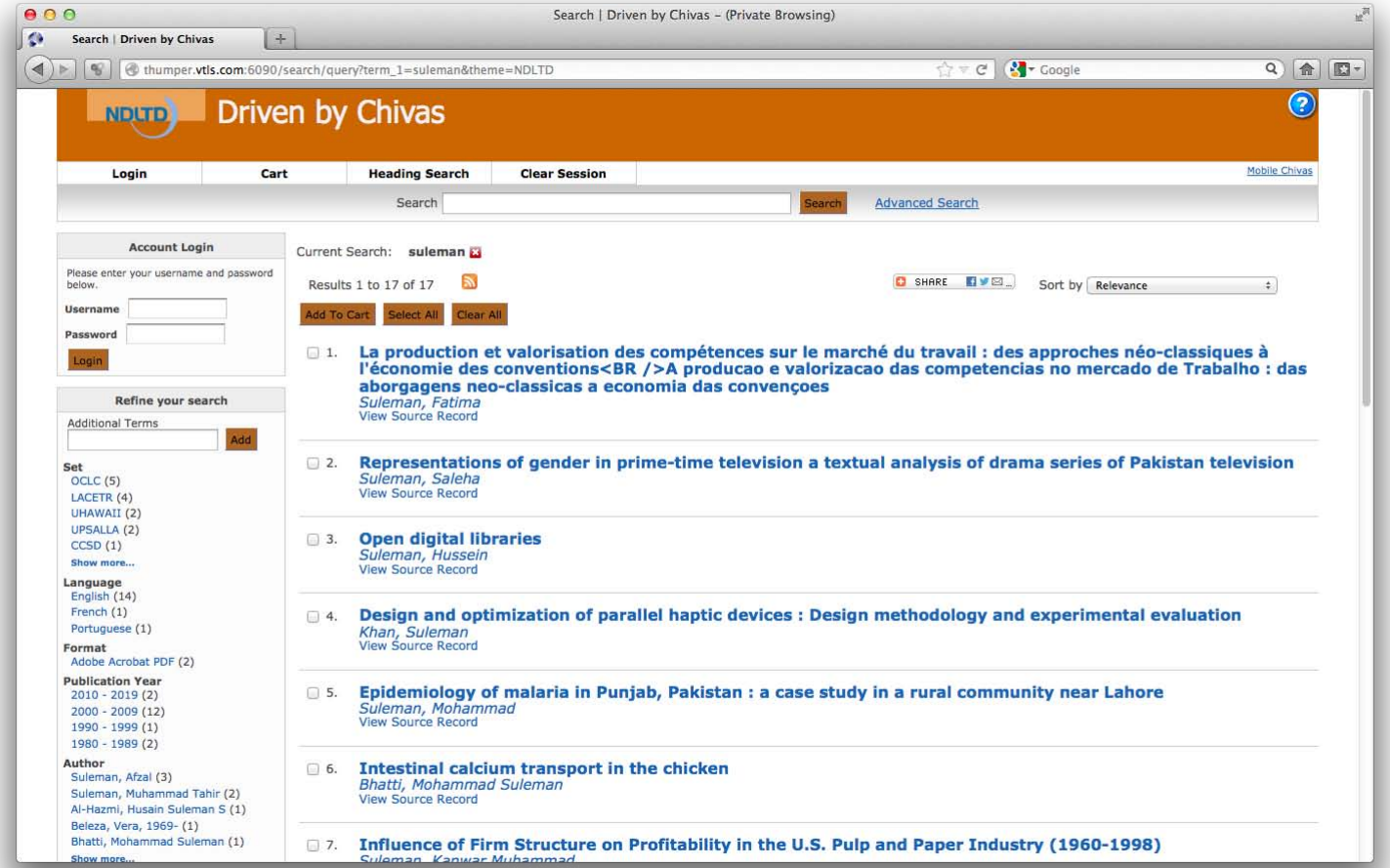
Recent Submissions

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

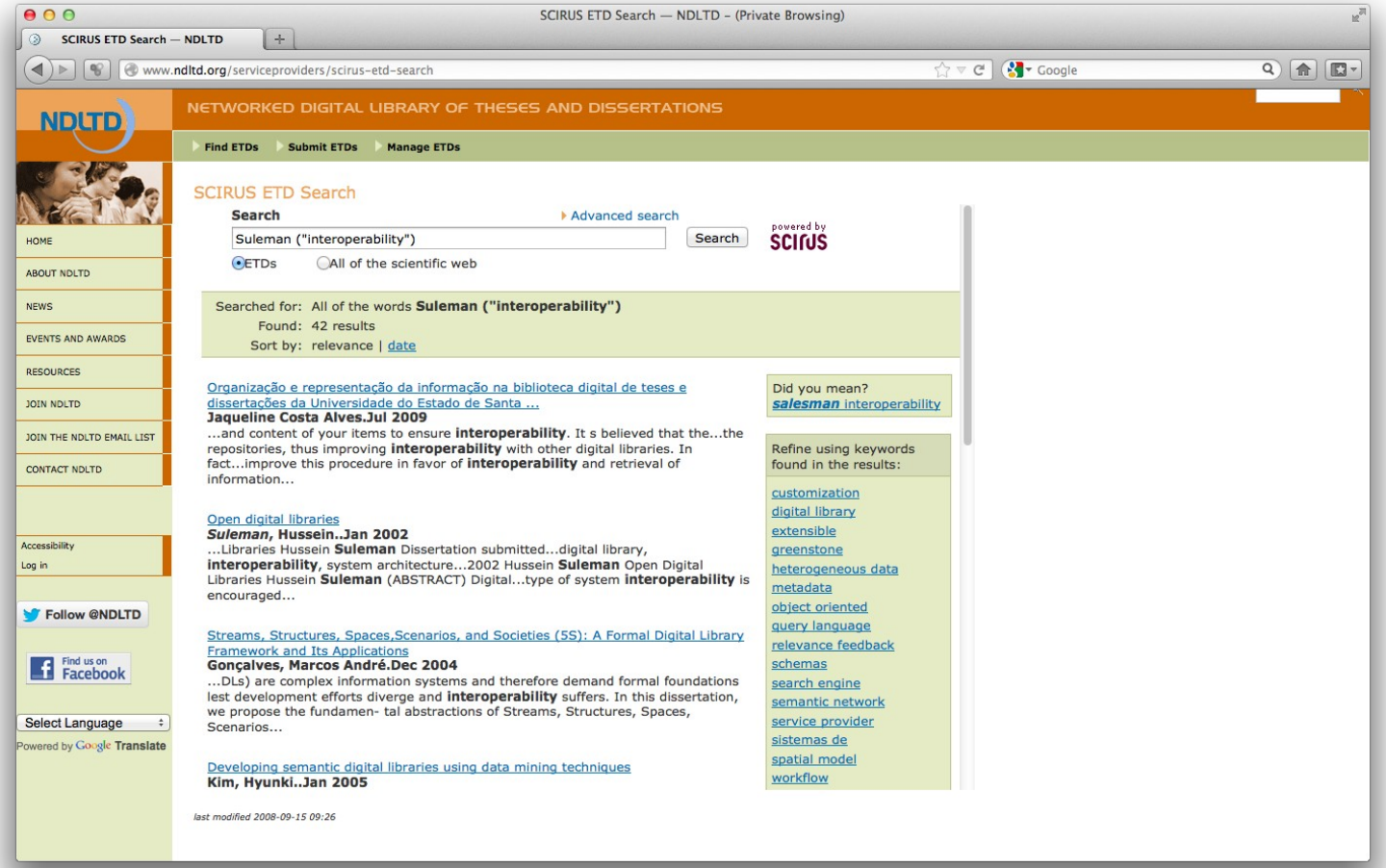
Collection Statistics

Collection	Total
CRANFIELD	22
Australiasian 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

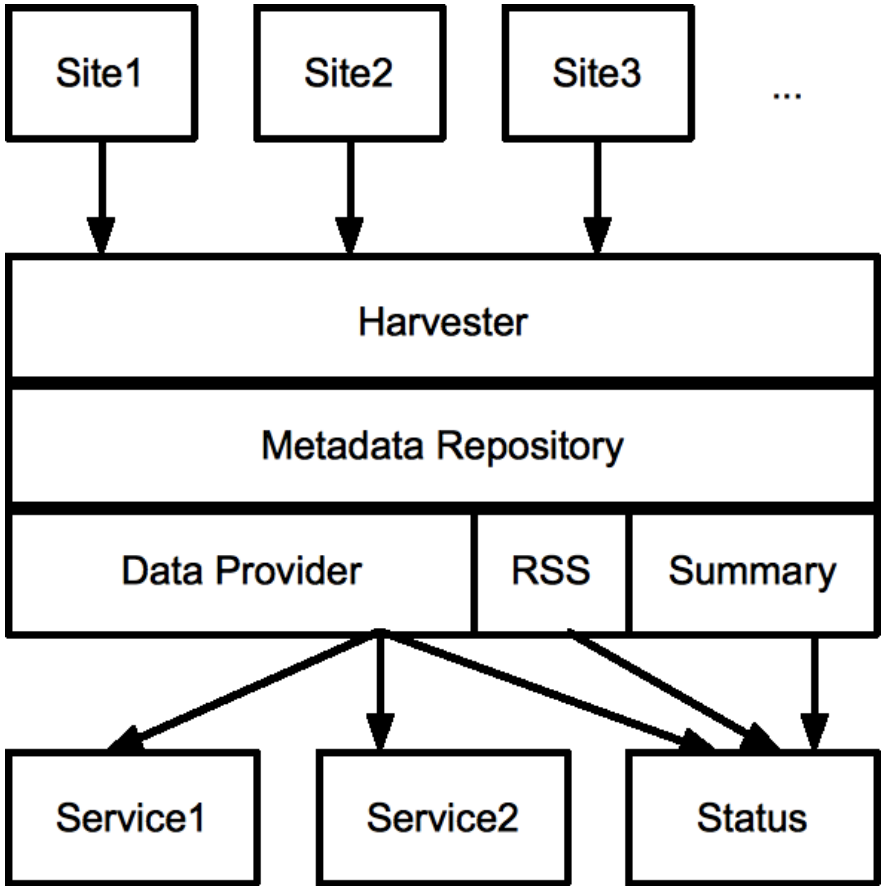
VTLS Discovery Service



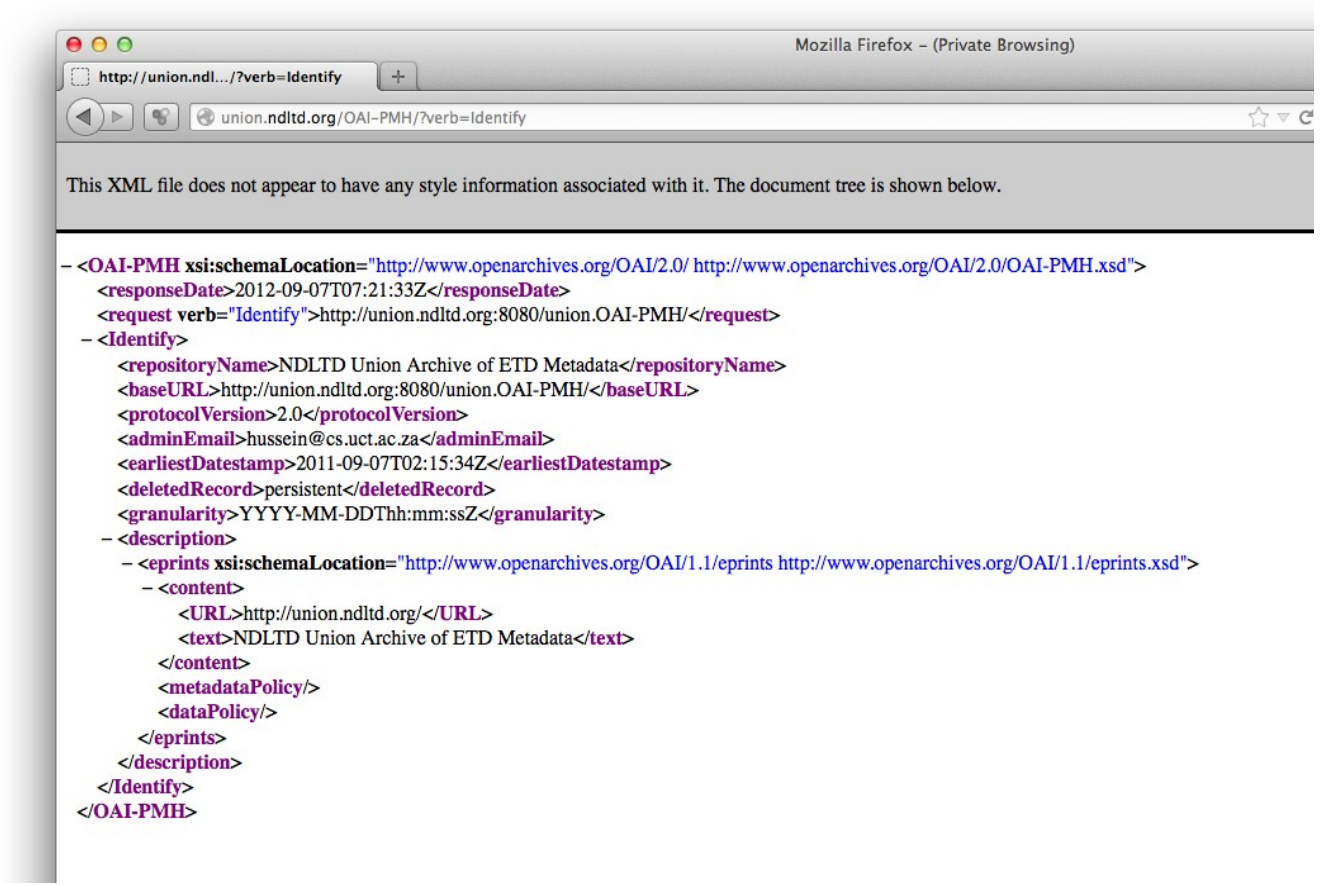
Scirus Discovery Service



How Does It All Work?

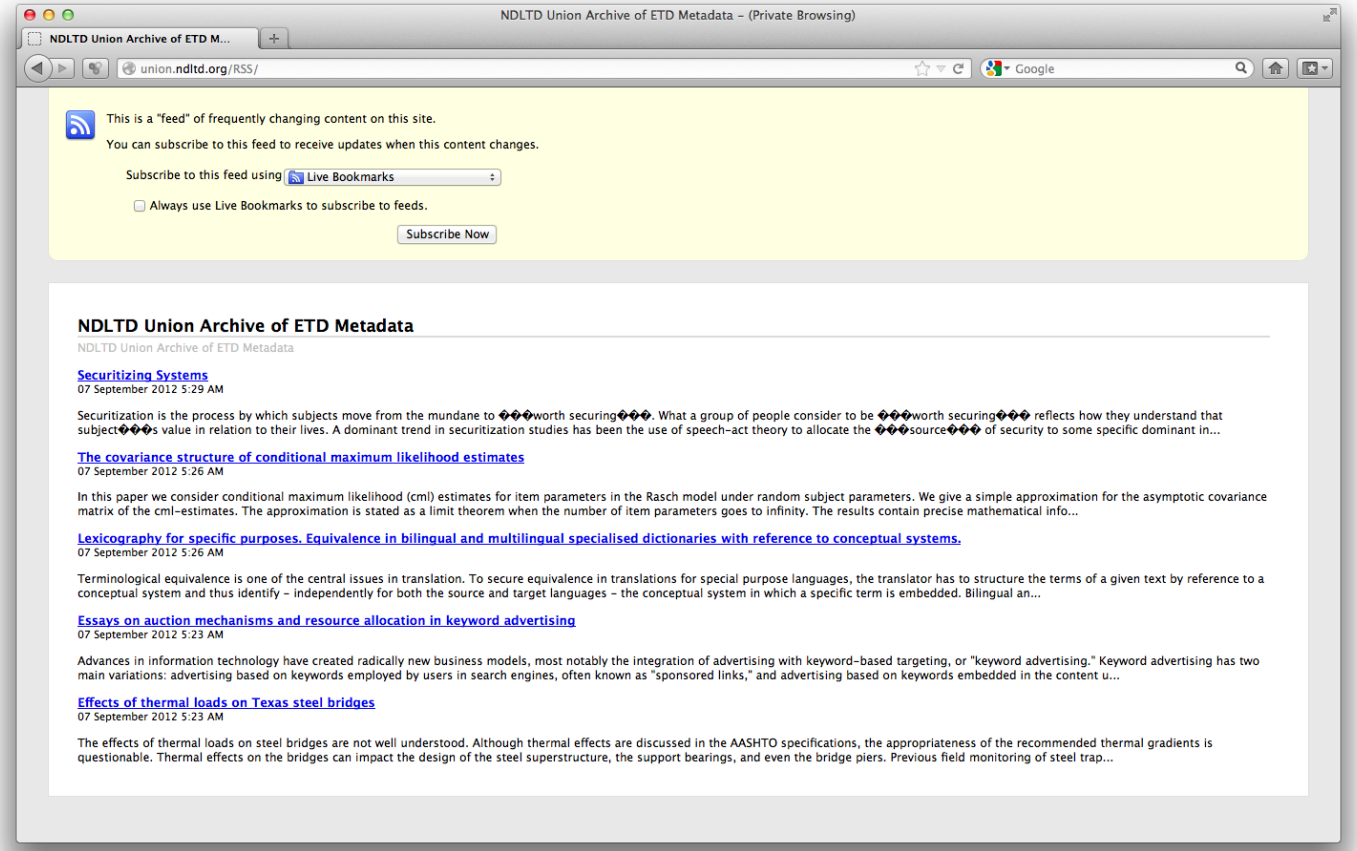


OAI-PMH?



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

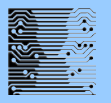
RSS?



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

Quick Numbers

- 130 remote sites
- 2 metadata formats: DC, ETDMS
- 1986187 records
- largest sites:
  - OCLC >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 ~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 exclusivly 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