Infrastructure & Harvesting Update

Tim Cole (t-cole3@illinois.edu)
University of Illinois at Urbana-Champaign

May 14, 2010
IMLS DCC Advisory Board Meeting
Harvesting

• Exclusively OAI-PMH (but testing other, more later)
  – Mostly oai_dc, a few qualified dc, one MARC, Aquifer MODS transformed to oai_dc
  – Metadata quality & consistency remains an issue

• Frequency ~ 4 harvests / repository / year

• Repositories / collections:
  – ~ 120 Aquifer repositories + 65 OH repositories
  – multiple “collections” per repository common
  – Over 1 million items in aggregate across portals

• Harvest problems / repeated harvests > 10%
post-harvest processing

• XML files filtered & processed before indexing
  – Often poor match between OAI set & collections

• Most pre-index processing via XSL-T
  – Add provenance & collection associations
  – Normalize type, date, spatial coverage
  – Too many collections now to specialize processing
  – Well formed XML fragments created for searching, faceting & browsing, presentation

• Driver upgraded from VBScript to C#.NET
  – Significant reductions in end-to-end process time
portal infrastructure

• Multiple portals
  – IMLS, Opening History, Transportation History
  – Developmental & rebuild portals

• Microsoft SQL Server data store, full-text index
  – Relational DB with XML columns, not fully normalized
  – Separate collection registry

• Transitioning from ASP to .NET + Web Services
  – New tabbed display heavily AJAX (asynchronous)
  – SQL & Web services separated
Issues

• Harvest quality & resources required
  – Still too many failed harvests; dead-end links in metadata
  – Normalization of heterogeneous metadata expensive & difficult; hard to verify accuracy & benefit
  – With growth, less able to customize process by provider

• SQL query complexity & correctness
  – Commercial SQL offers flexibility & control, but at cost of complexity and potential for error; not optimized for XML

• Performance & nimbleness
  – Too many dependencies on pre-index processing
  – Scaling concerns for facets, browse, more-like-this features; heavily spidered (persistent record ids, could use site map)
  – Even with AJAX/.NET, not as rapid to prototype as would like
Metasearch in IMLS DCC

William H. Mischo, Josh Bishoff
University of Illinois at Urbana-Champaign

May 14, 2010
IMLS DCC Advisory Board Meeting
University of Illinois Library Gateway

• Gateway Portal introduced in September 2007
  o Guide users to appropriate information resources
  o Recommender system
  o Integration of resources
  o Metasearch over 70 targets
  o Help with search strategy formulation and refinement

• Powered by custom metasearch system suite (Easy Search) that has been adapted in several settings.

• Approach centered around search strategy and reformulation assistance.
Easy Search Features

- Custom transaction logs: 4 million user search arguments; 3+ million clickthroughs
- Analysis of search arguments, pattern checking
- Result displays influenced by search arguments
- AJAX driven display
- Links into the native interfaces at the point of completed search
- NISO MXG support
Easy Search Functions

- Stopword removal -- implicit Boolean insertion
- Spelling suggestions
- Suggestions to limit to phrase and title word and phrase searches where appropriate
- Pattern matching for author search prompts
- Direct links to journal title matches
- Pre-Search lookup (and links) for frequently entered terms, pathfinder topics
- Pattern matching for link to Journal & Article Locator (full text article finder)
- Context sensitive arrangement of results
- Dark target searches in background
# UIUC Library Search Assistant Results

Search Term(s) Entered: stone circles

## Finished Searching

### Multi-Topic -- Journal and Magazine Articles

<table>
<thead>
<tr>
<th>Source</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Search Premier (Ebsco)</td>
<td>205 Article Matches</td>
</tr>
<tr>
<td>Academic Onefile (InfoTrac)</td>
<td>42 Article Matches</td>
</tr>
<tr>
<td>Scopus</td>
<td>337 Article Matches</td>
</tr>
<tr>
<td>Web of Science [Social Sci, Sciences, Medicine, Humanities, Engineering]</td>
<td>69 Article Matches</td>
</tr>
</tbody>
</table>

### Arts and Humanities

<table>
<thead>
<tr>
<th>Source</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLA International Bibliography</td>
<td>5 Article Matches</td>
</tr>
<tr>
<td>Historical Abstracts</td>
<td>7 Article Matches</td>
</tr>
<tr>
<td>America: History and Life</td>
<td>26 Article Matches</td>
</tr>
<tr>
<td>RILM Abstracts of Music Literature</td>
<td>1 Article Matches</td>
</tr>
</tbody>
</table>

### Education

<table>
<thead>
<tr>
<th>Source</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Full-Text</td>
<td>3 Article Matches</td>
</tr>
<tr>
<td>ERIC Education Literature (CSA)</td>
<td>2 Article Matches</td>
</tr>
<tr>
<td>Physical Education Index</td>
<td>No Matches on Search</td>
</tr>
</tbody>
</table>

### Web Search Engines

<table>
<thead>
<tr>
<th>Engine</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>21600000 Page Matches</td>
</tr>
<tr>
<td>Microsoft Live Web</td>
<td>33700000 Page Matches</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>400000 Page Matches</td>
</tr>
<tr>
<td>Scirus Academic Web Search</td>
<td>141202 Page Matches</td>
</tr>
<tr>
<td>referer</td>
<td>sessionid</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>http://...</td>
<td>607357558</td>
</tr>
<tr>
<td>http://...</td>
<td>607357558</td>
</tr>
<tr>
<td>http://...</td>
<td>607357576</td>
</tr>
<tr>
<td>http://...</td>
<td>607357576</td>
</tr>
<tr>
<td>http://...</td>
<td>607357595</td>
</tr>
<tr>
<td>http://...</td>
<td>607357595</td>
</tr>
<tr>
<td>http://...</td>
<td>607357633</td>
</tr>
<tr>
<td>http://...</td>
<td>607357615</td>
</tr>
<tr>
<td>http://...</td>
<td>607357628</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>searchstring</th>
<th>searchterm</th>
<th>catid</th>
<th>suggest</th>
</tr>
</thead>
<tbody>
<tr>
<td>kwj.g.march;</td>
<td></td>
<td>899</td>
<td>gateway gen, opac, web</td>
</tr>
<tr>
<td>fromauthclick:ln.march;fnj;miig.;</td>
<td></td>
<td>900</td>
<td>gateway gen, opac, web</td>
</tr>
</tbody>
</table>

| http://... | 607357648 | NULL | kw.PRK; |            | 947   | gateway gen, opac, web | Spell Found: the reductive pentose phosphate cycle; |
| http://... | 607357700 | NULL | kw.xin wenxue shilao |            | 948   | gateway gen, opac, web | |
| http://... | 607357634 | NULL | ti/fromresult: The reductive pentose phosphate cycle; |            | 949   | gateway gen, opac, web | Spell Found: phosphoribokinase and ribulose diphosphate carboxylase; |
| http://... | 607357634 | NULL | ti/fromresult:Phosphoribokinase and ribulose diphosphate carboxylase |            | 950   | gateway gen, opac, web | Insdpark pubmeddark |
| http://... | 607357714 | NULL | kw.performance of dowel bat refit projects in |            | 951   | gateway gen, opac, web | Insdpark |
| http://... | 607357714 | NULL | kw:field performance monitoring of repair treatment in |            | 952   | gateway gen, opac, web | Insdpark |
| http://... | 607357725 | NULL | kw:architettura processuale; |            | 953   | gateway gen, opac, web | |
| http://... | 607357731 | NULL | au/fromresult:The NAD-dependent glutamate dehydrogenase |            | 954   | gateway gen, opac, web | |
| http://... | 607357731 | NULL | au/fromresult:Ln.fromresult:weck in/fromresult:miig.; |            | 955   | gateway gen, opac, web | |
| http://... | 607357634 | NULL | ti/fromresult:The NAD-dependent glutamate dehydrogenase |            | 956   | gateway gen, opac, web | |
| http://... | 607357634 | NULL | ti/fromresult:Detection of glutamate dehydrogenase; |            | 957   | gateway gen, opac, web | |
| http://... | 607357634 | NULL | ti/fromresult:The NAD-dependent glutamate dehydrogenase |            | 958   | gateway gen, opac, web | pubmeddark |
User Studies

• Web searching behaviors
  o Short search sessions (72% are single query)
  o Average words per search query: 2.3 – 2.5
  o “Advanced features” not utilized or not utilized correctly
  o Users typically look at first page of results only

• Search Assistance
  • Users not using advanced search features, so we provide search suggestion prompts to provide powerful functionality
<table>
<thead>
<tr>
<th>Terms</th>
<th>Number of Queries</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18,845</td>
<td>11.4%</td>
</tr>
<tr>
<td>2</td>
<td>39,267</td>
<td>23.8%</td>
</tr>
<tr>
<td>3</td>
<td>39,095</td>
<td>23.7%</td>
</tr>
<tr>
<td>4</td>
<td>22,086</td>
<td>13.4%</td>
</tr>
<tr>
<td>5</td>
<td>13,814</td>
<td>8.4%</td>
</tr>
<tr>
<td>6</td>
<td>8,646</td>
<td>5.2%</td>
</tr>
<tr>
<td>7</td>
<td>5,506</td>
<td>3.3%</td>
</tr>
<tr>
<td>8</td>
<td>3,823</td>
<td>2.3%</td>
</tr>
<tr>
<td>&gt; 8</td>
<td>14,076</td>
<td>8.5%</td>
</tr>
<tr>
<td>Total:</td>
<td>165158</td>
<td>100%</td>
</tr>
</tbody>
</table>
Easy Search Transaction Log Analysis

49.4% of the searches were “known-item” as opposed to topical searches.

Of the 49.4% specific item searches:
- 7.4% of the 49.4% were author/title;
- 28.9% were author;
- 40.5% were book/monographic title searches;
- 6.8% were index/abstract title;
- 5.7% were for specific journal article; and
- 11.8% were for specific journal title

Overall, 17.96% of the searches contained a name or an organization, although clearly some of these are topical searches.
Search Behaviors
Fall 2009 Sample Searches (697,504)

• 12.17% use Boolean operators (11.95% are AND, 0.18% OR, 0.1% NOT) – but almost all are “and” within title or other phrase

• 6.64% use Commas

• 0.89% use Parentheses

• 3.81% use Quotes

• 22.47% use Prepositions (useful for parsing out phrases and identifying multiple concepts)

• 0.38% contain “+”
Search Sessions
Fall 2009 Sample Searches (697,504)

• Total of 286,748 search sessions; with an average of 2.43 searches per session
  • 50.15% of the sessions are single query (lower than web search studies)
  • 49.85% are 2 or more queries
  • 27.4% are 3 or more queries
  • 17.17% are 4 or more queries

• In 22.98% of sessions with more than one query, users clicked on a search assistance suggestion.
• The total percent of sessions where a user has clicked on a search assistance suggestion (22.98%) or performed a search reformulation from a pull-down menu (27.58%) is then 46.74%.
What We Have Learned

• Large numbers of specific (known) item searches being performed

• Users expect sophisticated parsing of their search terms and automatic advanced processing (e.g. truncation)

• Spell suggestions important

• Author and name searches must be parsed

• System is used as a Reference tool

• Search assistance in the form of system supplied strategic suggestions and reformulated tactical searching is being utilized in almost half of the search sessions of more than one query.
Implications for IMLS DCC

• Reformulation important

• New approach of providing secondary literature links using the primary source title and subject.

• System is used as a Reference tool

• Extensions to NISO MXG will allow additional reformulation and assistance
New Metasearch Mockups

Josh Bishoff, Yuting Li, Tim Cole, Katrina Fenlon
University of Illinois at Urbana-Champaign

May 14, 2010
IMLS DCC Advisory Board Meeting
Goal: find a way to direct users to secondary resources related to found items

• Entered search string is augmented with title and/or subject keywords from found records

• Marries “more like this” with standard metasearch techniques

• Evidence of potential, but challenging interface issues; results will not always be useful
Opening History: Transportation

Title
Pittsburgh and Lake Erie Repair Facilities

Identifier
http://images.library.pitt.edu/cgi-bin/image功效?view=entry&subview=details&imgid=K 1223 4580345 RR view id=4580345 RR TIF External Link

Publisher
Archives Service Center

Format
image/jpeg

Rights
http://www.library.pitt.edu/libraries/archives/police/guidelines.html

Type
Image

Date
1889-12-07

Description

Subject
McKees Rocks (Pa.); Pittsburgh and Lake Erie Railroad Company; Railroads--Pennsylvania--History; Locomotive works--Pennsylvania--McKees Rocks; Diesel locomotives--Pennsylvania--McKees Rocks.

Language
Undetermined

Creator
Pittsburgh and Lake Erie Railroad

Source
Pittsburgh and Lake Erie Railroad Company (ARC)

This item is part of the collection Historic Pittsburgh Image Collections

The Historic Pittsburgh Image Collections website contains over 8,000 visual images of the Pittsburgh and Western Pennsylvania region selected from dozens of photographic collections held by three cultural heritage institutions in Pittsburgh. The image collections visually document the cultural, educational, and social development of the city of Pittsburgh, as well as depict the vast infrastructure and industry of the region. Many portray Pittsburgh's large immigrant workforce, its steel industries, its public schools, its civic "renaissance" of the mid-1930s and 1930s, and its African-American communities.

View full collection details
View all 9323 items from this collection
Search for articles related to "Pittsburgh and Lake Erie Repair Facilities in America: History & Life" with your original query:

railroad accident

&
these terms from title:
- Pittsburgh
- Lake
- Erie
- Repair
- Facilities

&
these terms from subject:
- Railroads--Pennsylvania--History

Additional terms:

Search

Powered by the Grainger Engineering Library
Search for articles related to
"Pittsburgh and Lake Erie Repair Facilities"
in America: History & Life

with your original query:
railroad accident

\&
these terms from title:
- Pittsburgh
- Lake
- Erie
- Repair
- Facilities

\&
these terms from subject:
Pennsylvania--History

Additional terms:

Search

Powered by the Grainger Engineering Library
Results: 1-3 of 3

1. Public Relations and Technology: The "Standard Railroad of the World" and the Crisis in Railroad Safety, 1867-1916
   Subjects: Pennsylvania; Railroads; Railroad accidents -- Law & legislation; Industrial safety -- Government policy; Transportation -- Safety measures; Government regulation; Industrial safety; Pennsylvania Railroad Co.; Public relations; Technology
   Database: America: History & Life

2. A Bridge Builder Changes a Railroad: The Story of Daniel Craig McCallum
   Subjects: New York (State); Pennsylvania; Railroads; Organizational structure
   Database: America: History & Life

3. The Great Locomotive Wreck
   By: Jackson, Jack. Civil War Times Illustrated, Jan/Feb 95, Vol. 33 Issue 1, p48, 6p, 4 Color Photographs, 2 Black and White Photographs; Historical Period: 1861 to 1865; (AM 950119405)
   Subjects: United States -- History -- Civil War, 1861-1865 -- Transportation; Railroad accidents -- History; Railroads; Prisoners of war; Civil war; Accidents
   Database: America: History & Life
Supplementing OAI-PMH Harvesting in the IMLS Digital Collections & Content Repository

Jacob Jett (jjett2@illinois.edu)
University of Illinois at Urbana-Champaign
Summary of the contents of the IMLS DCC Registry and Repository (as of 12/15/2009)

- Collections in the Registry: 341
- Items in the Repository: 630,619
- Collections from which items have been harvested via OAI-PMH: 78
- Collections from which no items have been harvested: 263
Known barriers to interoperability, as identified by Sarah Shreves in 2005.

- Insufficient technical infrastructure to make implementation of OAI applications practicable.
- Insufficient metadata or metadata that was too poorly formed to make sharing practicable.
- Socio-economic factors specific to each individual institution and project.
Research Goals

• Reassess the 25 largest collections that have had no items harvested to:
  – Determine if they have implemented OAI-PMH and can be harvested at this time.
  – Determine if item-level records can be harvested from them using means other than OAI-PMH, if the latter is not available.
  – Determine if item-level access can be achieved through the use of metasearch technology if harvesting item-level records proves to be impractical.

• Experiment with http technologies to:
  – Harvest item-level metadata from the candidate collections.
  – Retrieve item-level metadata from the candidate collections’ native search interfaces.

• Compare and contrast the value of these two supplemental approaches to metadata access in the context of the DCC project.
Initial Findings – Collections that have implemented OAI-PMH and that can now be harvested.

- Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois (aka: Illinois Historical Aerial Photography 1938-1941)
- George Edward Anderson Collection
- HEBRTH (Home Economics Archive: Research, Tradition, and History)
- Montana Memory Project
Initial Harvesting Results

• 122,274 item-level records harvested from 5 collections

• Collections can be broadly classified into 2 types according to their interface structure:
  – Heterogeneous home-grown interfaces
  – Homogeneous off-the-shelf interface applications (e.g. CONTENTdm)
Sample web page for harvesting.

Thanksgiving Day at Casco Street School
Contributed by Maine Historical Society/MaineToday Media

Description
Three children re-enacting the Pilgrim's Thanksgiving Feast during a food drive at the Casco Street School in Portland, 1922.

Other Information
Title: Thanksgiving Day at Casco Street School
Creator: Portland Press Herald
Creation Date: 1922
Subject Date: 1922

Purchase a reproduction of this item on VintageMaineImages.com.
Sample HTML ingested by the harvester.

- 
- 
- 
- 
  <h3>Description</h3>  
  <p>Three children re-enacting the Pilgrim’s Thanksgiving Feast during a food drive at the Casco Street School in Portland, 1922.</p>  
- 
  <h3>Other Information</h3>  
  <ul class="nomark noindent">  
    <li><strong>Title:</strong> Thanksgiving Day at Casco Street School</li>  
    <li><strong>Creator:</strong> Portland Press Herald</li>  
    <li><strong>Creation Date:</strong> 1922</li>  
  </ul>  
- 
- 
-
Thanksgiving Day at Casco Street School

Portland Press Herald

Community Service -- Maine -- Portland

Casco Street School (Portland, Me.)

Celebrations -- Maine -- Portland

Charity -- Maine -- Portland

Thanksgiving Day

Costumes -- American

Three children re-enacting the Pilgrim's Thanksgiving Feast during a food drive at the Casco Street School in Portland, 1922.
Preliminary Outcomes

• Generally, should be able to harvest from any data interface that provides either a set of indexed browsing pages or which uses an iterative identification number scheme as part of their resource’s uri’s.

• Will be likely to need metasearch techniques to access item-level descriptions for those data providers who do not have indexed browsing available via their data interfaces and who are using an record identification that is strictly internalized within the database.
Some Observations

• Collections can also be classified by:
  – Content type:
    • Simple surrogates, complex surrogates, register entries, etc.
  – Source type:
    • Digital Archive, Digital Collection, Digital Library, Digital Museum, Digital Registry, etc.

• Many opportunities to add value to content:
  – Added contextualization
  – Standard metadata scheme
  – Normalization?
Research is ongoing

• Continuing to experiment with CONTENTdm harvesting to verify the generalizability of harvest results.

• Expanding harvesting activities to other standard data storage/interface packages such as:
  – PastPerfect
  – ActivePaper

• Metasearch application to be built.

• Scalability questions yet to be answered.