IntegraL: Link-based Integration of Library Digital Collections and Resources

NJLA Conference: NJLA CUS/ACRL-NJ Technology Innovation Forum
April 27, 2010

Ann D. Hoang
Assistant University Librarian
New Jersey Institute of Technology
ahoang@njit.edu
973-596-5798
Members of the Team …

Min Song
Assistant Professor
Information Systems Dept. @ NJIT
http://is.njit.edu/people/song.php

IntegraL Project Manager and Co-Principal Investigator

Research interests:
- Text mining
- Dynamic document clustering
- Semantic query expansion
- Bioinformatics
- Information extraction & retrieval

Ann D. Hoang
Assistant University Librarian
Van Houten Library @ NJIT
http://web.njit.edu/~ahoang

IntegraL Library Technical Support and System Evaluation

Responsible for:
- Staff, services related to technology including web services, ILS system and digital technology
- Acquisitions of print collections & cataloging
- Improve quality of services across all library functions
Members of the Team ...

- **Michael Bieber**, Principal Investigator
- **Richard Sweeney**, NJIT University Librarian
- **Vahid Hamidullah**, Undergraduate Student
- **Eric Koppel**, Undergraduate Student
- **Pawel Bokota**, Undergraduate Student
- **Said Bleik**, Ph.D. Student
- **NJIT Librarians**
The Question …

Why do we need libraries, library resources or librarians when we have Google?

A typical question often asked by inexperienced library users … not by serious information seekers.
Invisible Web aka the “Deep Web”

**Robert Lackie** spent most of the last decade educating libraries users and librarians about “Those Dark Hiding Places: The Invisible Web …”
http://www.robertlackie.com/

**Michael Bergman**, a web scientist, first coined the phrase “Deep Web”
http://www.mkbergman.com/

Deep Web refers to web content that are created dynamically as a result of a specific search executed by a searcher.
The Deep Web ...

- “Serious information seekers can [not] avoid the importance or quality of deep Web information.”
- “Deep Web information is only a component of total information available.”
- “Searching must evolve to encompass the complete Web.”
Exploring a ‘Deep Web’ That Google Can’t Grasp

By ALEX WRIGHT

One day last summer, Google’s search engine trundled the list of Web pages it knows about. But as impossible as it may seem, it’s impossible to know the entire Web.

Beyond those trillion pages lies an even vaster Web of information—raw data, medical records, financial schedules, medical research and all kinds of other materials stored in databases that remain invisible to search engines.

The challenges that the major search engines face are immense, explaining why they still can’t provide satisfactory answers to questions like: what is the capital of Switzerland? Which is the capital of the United Kingdom? London next Thursday? The answers are all there, but the search engines can’t find them.

Now a new breed of technologies is taking us into those corners. When that happens, it will do more than merely provide an answer. It will change the way many companies do business online.

"...added the one trillionth address to the list of web pages... beyond those trillion pages lies... web of hidden data... all kinds... materials stored in databases that remain... invisible to search engines."

"...deep web technologies hold the promise of achieving [web data integration], by automating the process of analyzing database structures and cross-referencing the results."
The Web + The User…

• The Web is …
  – incredibly successful at delivering information to users and searchers;
  – not yet a well organized repository of presentable structured HTML pages;
  – a conglomerate of volatile HTML pages.

• The User expects …
  – an urgent need to go beyond browsing;
  – response dynamically to users’ queries;
  – wants to get there faster.

Web of Organized Repository of Structured Content

… health site could cross-reference data from pharmaceutical companies with the latest findings from medical researchers, or

… a local news site could extend its coverage by letting users tap into public records stored in government databases.
You mean all that relevant information I need is hidden. I’m unable to access it and therefore preventing me from getting an “A” on my research paper? What will I tell my parents?

Challenges for searchers

- need to know what resources to use before they can access them,
- find related information from other resources,
- leave current resource to search related information from other resources.
What is IntegraL?

Fear not, information seekers. IntegraL is here to save you from a lifetime of drudgery. The “A” that you are expecting is in your future.

Background

- research funded by National Science Foundation and Institute of Museum and Library Services;
  - develop a system to integrate library resources to make searching easier and faster;
  - bring relevant and related content directly to the users;
  - improve the effectiveness of both users and library resources
What is IntegraL? ...

Step right up, folks! Witness amazing feats of bravura and derring-do! Watch as search results transition before your very eyes!

... is an open source software search engine
... is designed to allow users to efficiently search online resources
  – authenticated + non-authenticated
... is browsable across online resources through “linking” of element of interest icons
What is Integral? ...

Element of Interest

Icons highlight Element of Interest (EOIs)
- allow ease of integration between various resources
- supports drill-down for more detailed information on Elements of Interest (EOIs)
- empowers user’s deep search via NER (name entity recognition) plug-ins

1. Semantic deep web: automatic attribute extraction from the deep web data sources
   Yoo Jung An, James Geller, Yi-Ta Wu, Soon Ae Chun
   Publisher: ACM
   Request Permissions
   Full text available: Pdf (218.34 KB)

Additional Information: full citation, abstract, re
What is Integral? …

Link-based integration
What is IntegraL? …

IntegraL = Integrating + Libraries

- IntegraL search engine focuses on:
  - linking online resources selected by librarians
  - providing customized user preferences to the types of linked information.

- IntegraL system architecture:
  - Provide searchers with multiple access points
  - enables searchers to navigate unlinked web content
  - seamlessly links various resources to make browsing across resources easier
  - SSO with Shibboleth
    - your credentials is recognize across various services including email servers, database subscriptions, chat, etc. on your network.
System Architecture

IntegraL Search

Google
Locally managed collections
Online Catalog
Search Engines
Commercial database

Result Lists Display

IntegraL Federated Results Lists

Link-Based Integration
Selection of New Resources

Meta-information Engine: ME Lexical Analysis, ME Broker, ME Mapping Engine and Schema, Shibboleth, Plug-ins

Fulltext in databases
Personal Webpage
Images
Fulltext on surface & deep web
System Architecture ...

Administrative Panel allows librarians to make changes to IntegraL.

Add / Modify:
- Plugins,
- Engines (online resources)
- Categories (subjects)
- Elements of Interest and
- User Roles
All right, folks. Stop and relax. We are entering the demo zone.

IntegraL Demo
IntegraL Demo ...

Integrating Libraries Research

Project Overview

**Integrating Libraries Research (IntLib Project and IntegraL System)**

Mission Statement

IntegraL makes library users more effective by providing personalized, optimized library resources for their specific tasks. The IntegraL projects provide digital library resources nationwide. IntegraL provides services, collaborative filtering and federated metasearch.

The IntegraL Research Projects

NJIT’s Integrating Libraries research comprises three related, complementary research projects hosted at the New Jersey Institute of Technology (NJIT) - Digital Library Service Integration (DLSI), IntegraL and IntLib. We also cooperate closely with NJIT’s General Recommendation Engine (GRE) project.

http://base.njit.edu
IntegraL Demo

You only have to login once. Your system passes your credentials onto various services on the network.
IntegraL Demo

IntegraL Federated Search

Your Preferences  Digital Libraries  Logout

Search

IntegraL Integrating Libraries Research

Project Overview

Integrating Libraries Research (IntLib Project and IntegraL System)

Mission Statement

NJIT New Jersey's Science & Technology University

THE EDGE IN KNOWLEDGE
IntegraL Demo ...

Where EOI elements are selected

Resources defined by a librarian
IntegraL Demo ...

Your Preferences | Digital Libraries | Administration Panel

“michael bergman” “deep web”

Federated Search Results

<table>
<thead>
<tr>
<th>Source</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>4200</td>
</tr>
<tr>
<td>Ask Search</td>
<td>132</td>
</tr>
<tr>
<td>NJIT Library Author Search</td>
<td>0</td>
</tr>
<tr>
<td>NSDL Search</td>
<td>0</td>
</tr>
<tr>
<td>ACM Search</td>
<td>0</td>
</tr>
<tr>
<td>Google Book Search</td>
<td>0</td>
</tr>
<tr>
<td>GoogleImages</td>
<td>93</td>
</tr>
<tr>
<td>Scopus Affiliation Search</td>
<td>0</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>40</td>
</tr>
<tr>
<td>IEEE Book Search</td>
<td>0</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td>1</td>
</tr>
</tbody>
</table>
IntegraL Demo …

Google Scholar and Science Direct Results

Distributing non-MARC metadata: the CUGIR metadata sharing project
A Chandler, EL Westbrooks - Library Collections, Acquisitions and ..., 2002 - Elsevier

... The first problem was that individual metadata records are not accessible in CUGIR because they reside in the “Deep Web.” In the white paper “The Deep Web: Surfacing Hidden Value,” Michael Bergman of BrightPlanet Inc. defines Deep Web as being: ...
Distributing non-MARC metadata: the CUGIR metadata approach

Adam Chandler \textsuperscript{a} and Elaine L. Westbrook\textsuperscript{b}

\textsuperscript{a} Cornell University, Central Technical Services, 110 C. Olin Library, Ithaca, NY 14853
\textsuperscript{b} Cornell University, Albert R. Mann Library, Technical Services, Ithaca, NY 14853-5201

Available online 1 October 2002.

Brought to you by:
the NJIT Library
IntegraL Demo ...

<table>
<thead>
<tr>
<th>Federated Search Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google</strong></td>
</tr>
<tr>
<td><strong>Ask Search</strong></td>
</tr>
<tr>
<td><strong>NJIT Library Author Search</strong></td>
</tr>
<tr>
<td><strong>NSDL Search</strong></td>
</tr>
<tr>
<td><strong>ACM Search</strong></td>
</tr>
<tr>
<td><strong>Google Book Search</strong></td>
</tr>
<tr>
<td><strong>GoogleImages</strong></td>
</tr>
<tr>
<td><strong>Scopus Affiliation Search</strong></td>
</tr>
<tr>
<td><strong>Google Scholar</strong></td>
</tr>
<tr>
<td><strong>IEEE Book Search</strong></td>
</tr>
<tr>
<td><strong>ScienceDirect</strong></td>
</tr>
</tbody>
</table>
IntegraL Demo ...

Link-based integration on keywords for Database searching, IS and ACM Digital Library.
Link-based integration on keyword for Database searching, IS and ACM Digital Library
### IntegraL Demo ...

#### Federated Search Results

<table>
<thead>
<tr>
<th>Search Type</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>245000</td>
</tr>
<tr>
<td>Ask Search</td>
<td>41900</td>
</tr>
<tr>
<td>NJIT Library Author Search</td>
<td>5599</td>
</tr>
<tr>
<td>NSDL Search</td>
<td>7</td>
</tr>
<tr>
<td>ACM Search</td>
<td>442</td>
</tr>
<tr>
<td>Google Book Search</td>
<td>0</td>
</tr>
<tr>
<td>GoogleImages</td>
<td>15100</td>
</tr>
<tr>
<td>Scopus Affiliation Search</td>
<td>0</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>4280</td>
</tr>
<tr>
<td>IEEE Book Search</td>
<td>0</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td>163</td>
</tr>
</tbody>
</table>
Link-based integration

author: James Geller, General, Google
James Geller is affiliated with NJIT and his research interest is in semantic web.
Benefits of Using IntegraL

Browsing cross resources

- multiple access points through linking of EOIs,
- ability to navigate unlinked web content e.g. from propriety databases to author’s personal webpage,
- single sign-on authentication using Shibboleth, to seamlessly links to databases an resources that requires a password login.
- users awareness of additional resources

Logical? What the heck are these bozos talking about? Clowns don’t do logical.
Benefits of Using IntegraL …

- Direct access to system’s features
  - access to admin panel via the web
  - Ability add/modify/delete resources with minimal technical expertise.

- Information Literacy
  - IntegraL automates a search strategy that experienced searchers might go through by manually and inexperienced searchers might not even be aware how to do.

- Bridging silos between surface web and deep web
  - This bridges the gap between the silos of the structured web and the deep web
  - IntegraL leverages metadata and structure from one silo into the another to create a better search toward getting better results.
Links & Further Reading


- Bradley, B. (2007). Debate: Do we need librarians now we have Google? [http://www.slideshare.net/Philbradley/debate-do-we-need-librarians-now-we-have-google](http://www.slideshare.net/Philbradley/debate-do-we-need-librarians-now-we-have-google)


Thanks.

Questions?

Are we done yet?