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 …

NEWJERSEY Library Association

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 …

NEWJERSEY Library Association

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.**

The Answer …

NEWJERSEY Library Association

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

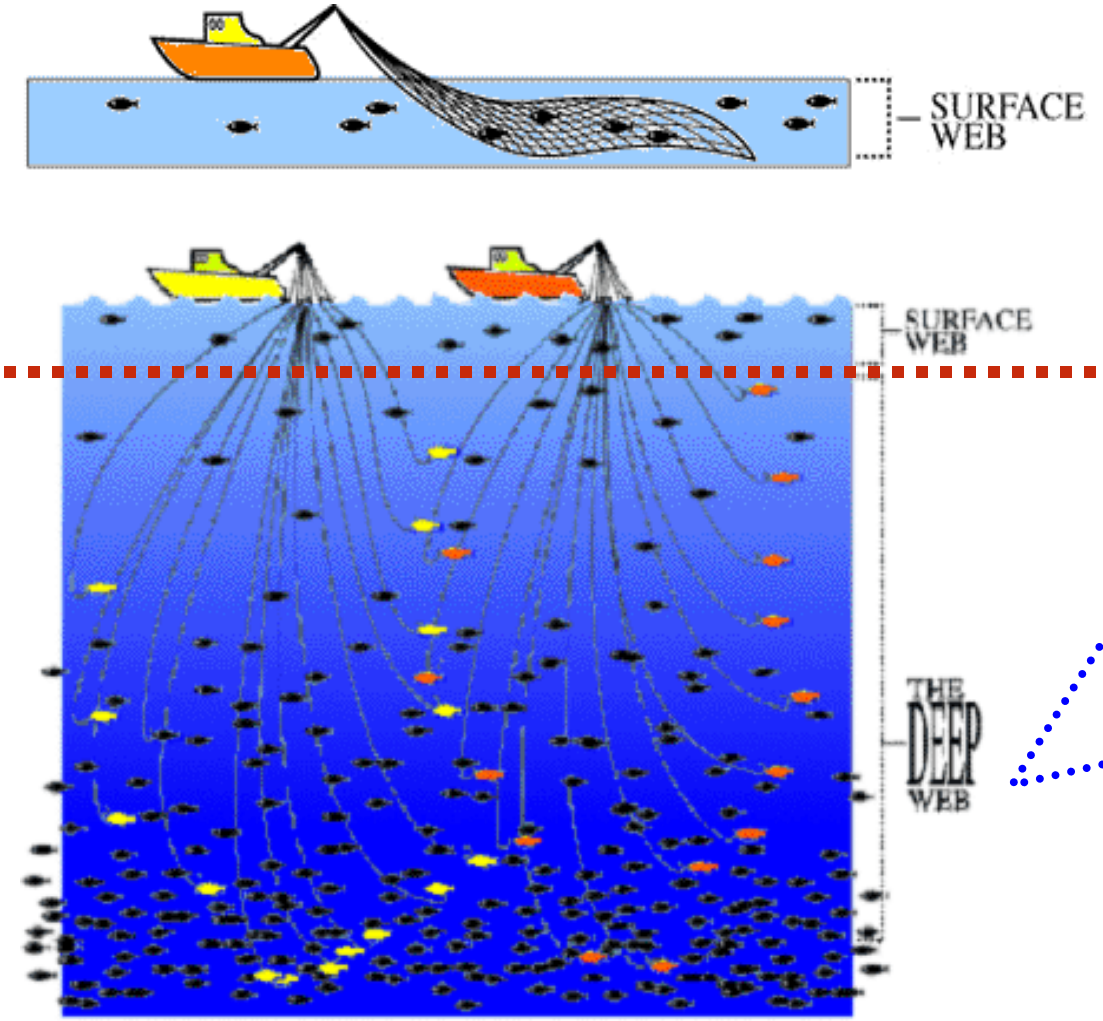
"**Deep Web**"

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 …

NEWJERSEY Library Association



- *"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."*

The Deep Web …

NEWJERSEY Library Association

Google does not bring relevant resources directly to the users.

The New York Times

This copy is for your personal, noncommercial use only. You can order presentation-ready copies for distribution to your colleagues, clients or customers here or use the "Reprints" tool that appears next to any article. Visit www.nytreprints.com for samples and additional information. Order a reprint of this article now.

February 23, 2009

Exploring a 'Deep Web' That Google Can't Grasp

By ALEX WRIGHT

One day last summer, Google's search engine trundle
the list of Web pages it knows about. But as impossibl
entire Web.

Beyond those trillion pages lies an even vaster Web o
schedules, medical research and all kinds of other mat
engines.

The challenges that the major search engi
explaining why they still can't provide sati
London next Thursday?" The answers are

Now a new breed of technologies is taking
corners. When that happens, it will do mo
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…

NEWJERSEY Library Association

- 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.

The Challenge …

NEWJERSEY Library Association

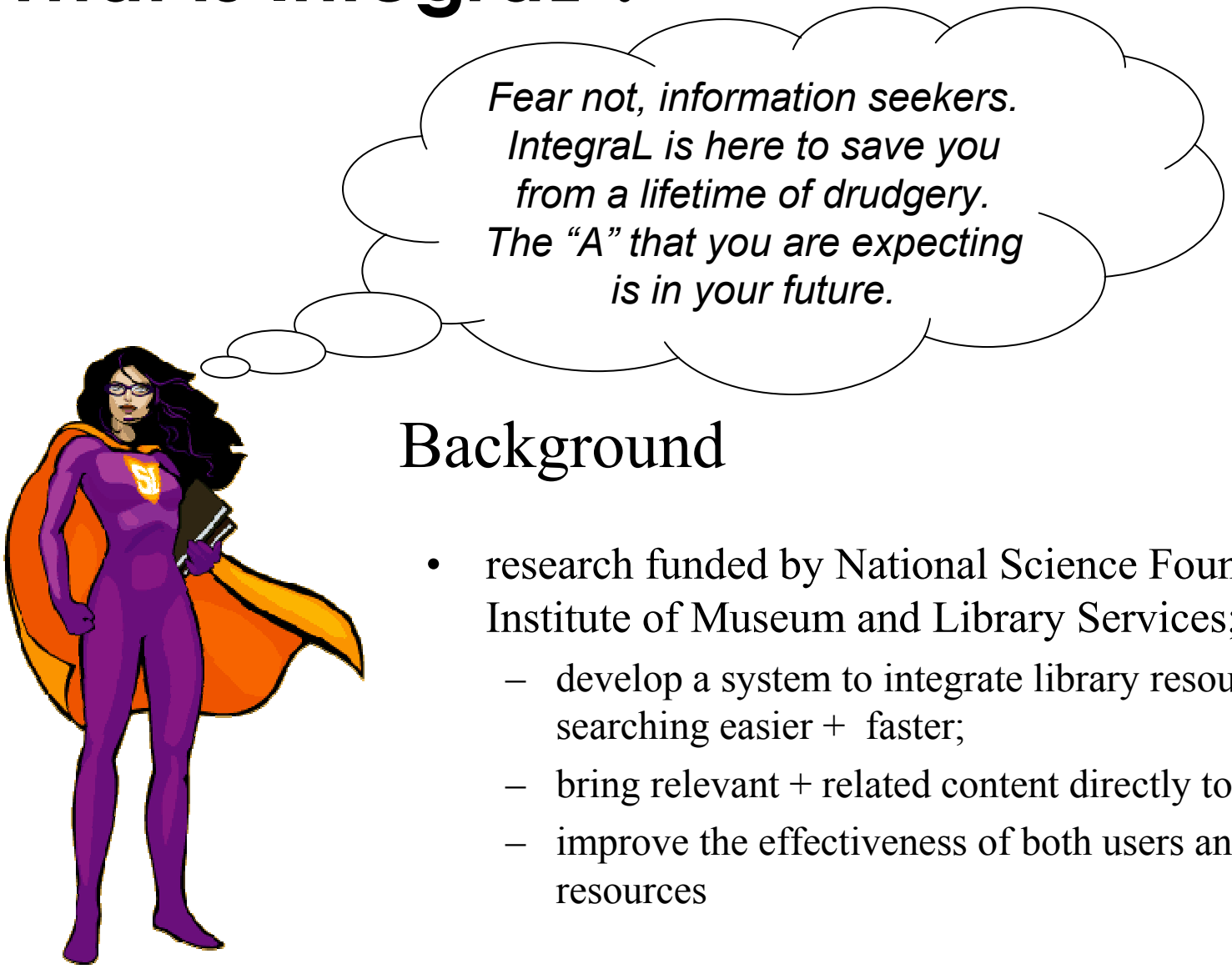
You mean all that relevant information I need is hidden. I'm unable to access it and therefore preventing me from getting an "A" on my research paper? What will I tell my parents?

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 ?

NEWJERSEY Library Association



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 + faster;
  - bring relevant + related content directly to the users;
  - improve the effectiveness of both users and library resources

What is IntegraL? …

NEWJERSEY Library Association

... 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? …

NEWJERSEY Library Association

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 if 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

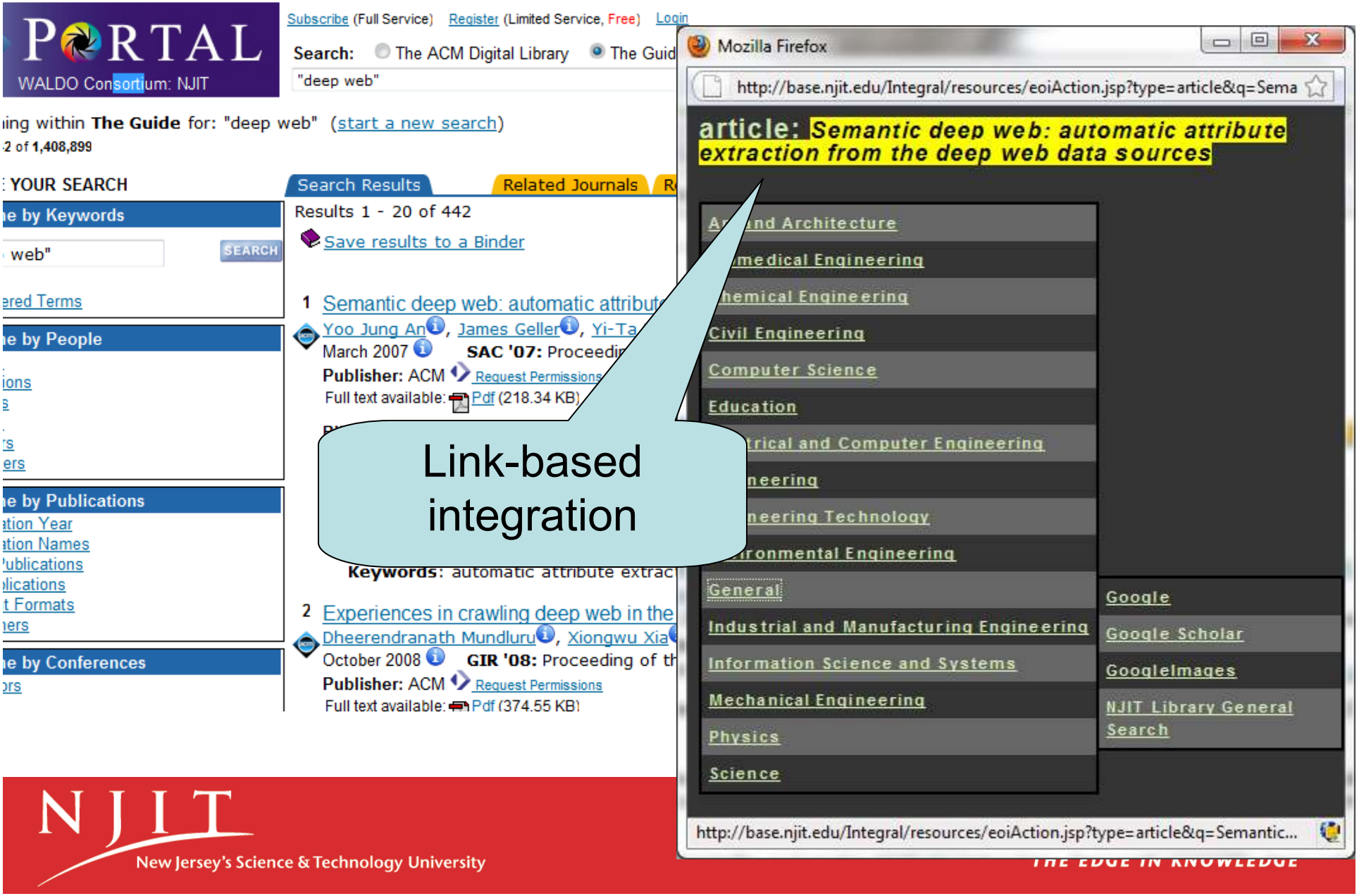
March 2007 **SAC '07:** Proceedings of the 2007 ACM symposium on Applied computing

Publisher: ACM Request Permissions

Full text available: Pdf (218.34 KB) Additional Information: full citation, abstract, re

What is IntegraL? …

NEWJERSEY Library Association



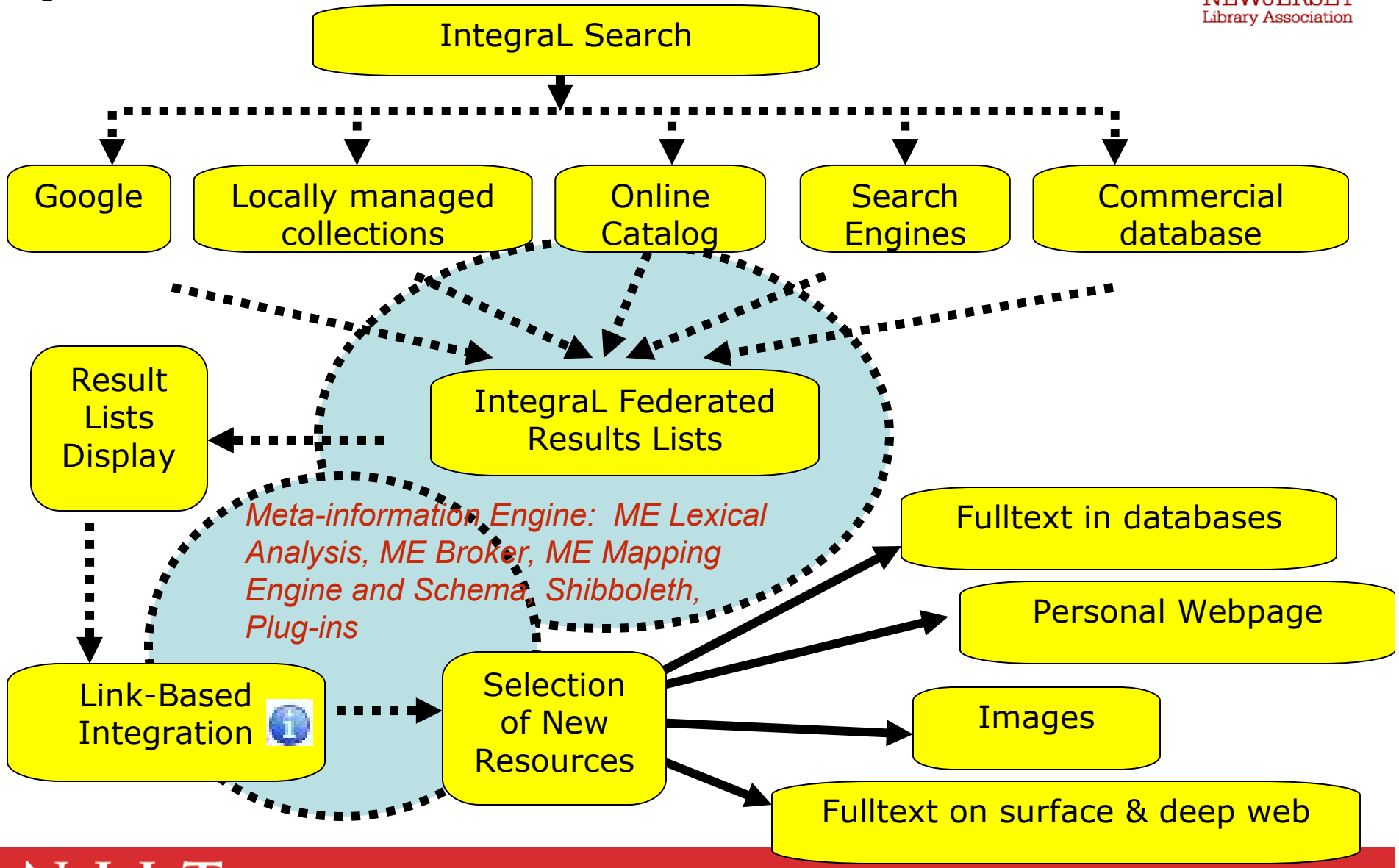
What is IntegraL ? …

NEWJERSEY Library Association

IntegraL = **Integra**ting + **L**ibraries

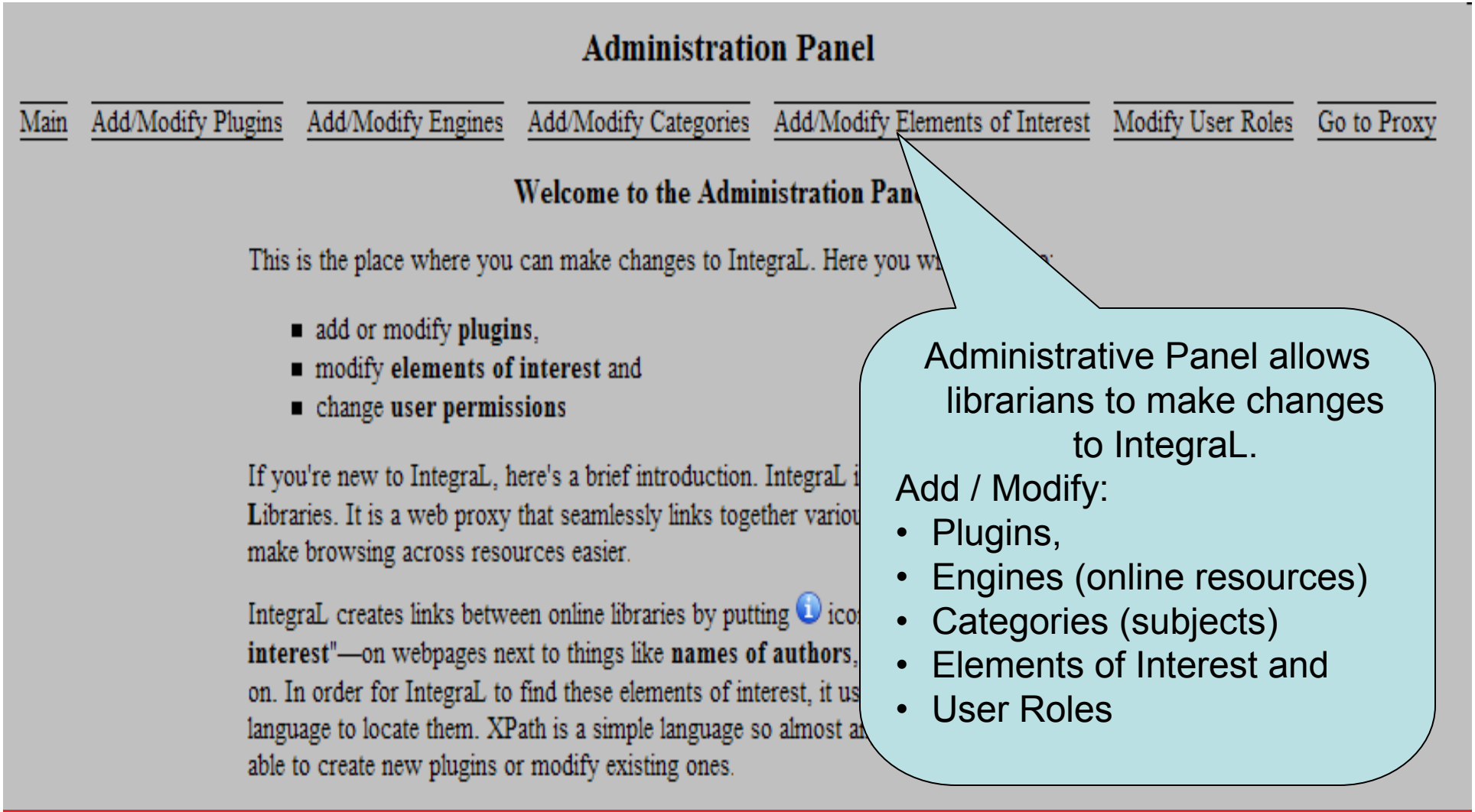
- IntegraL search engine focuses on:
  - linking online resources selected by librarians
  - providing customized user preferences to the types of linked information.
- IntegraL system architecture:
  - Provide searchers with multiple access points
  - enables searchers to navigate unlinked web content
  - 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



System Architecture …

NEWJERSEY Library Association



NEWJERSEY
Library Association

IntegraL Demo

IntegraL Demo …

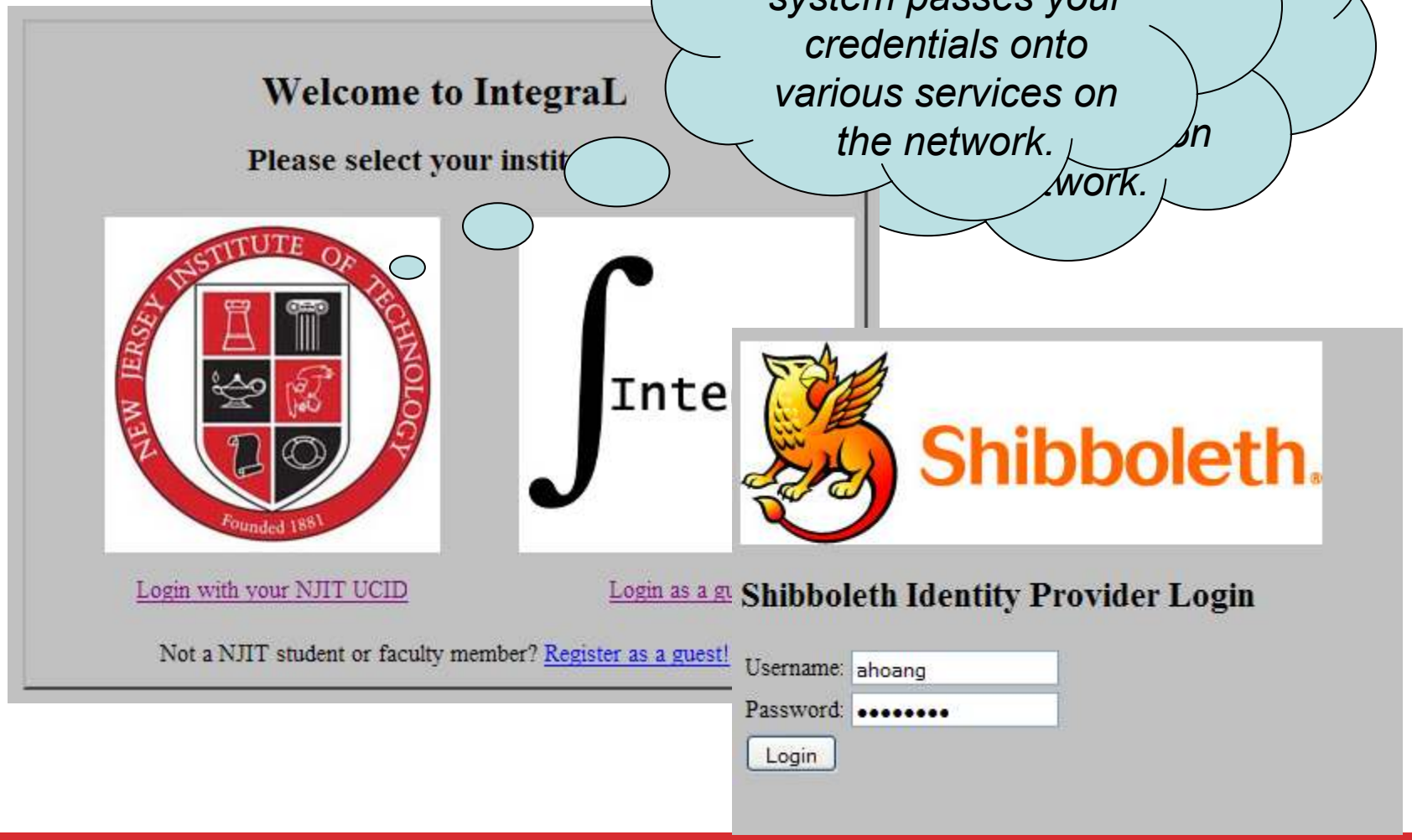
NEWJERSEY Library Association

NJIT New Jersey's Science & Technology University

IntegraL Demo

NEWJERSEY Library Association

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



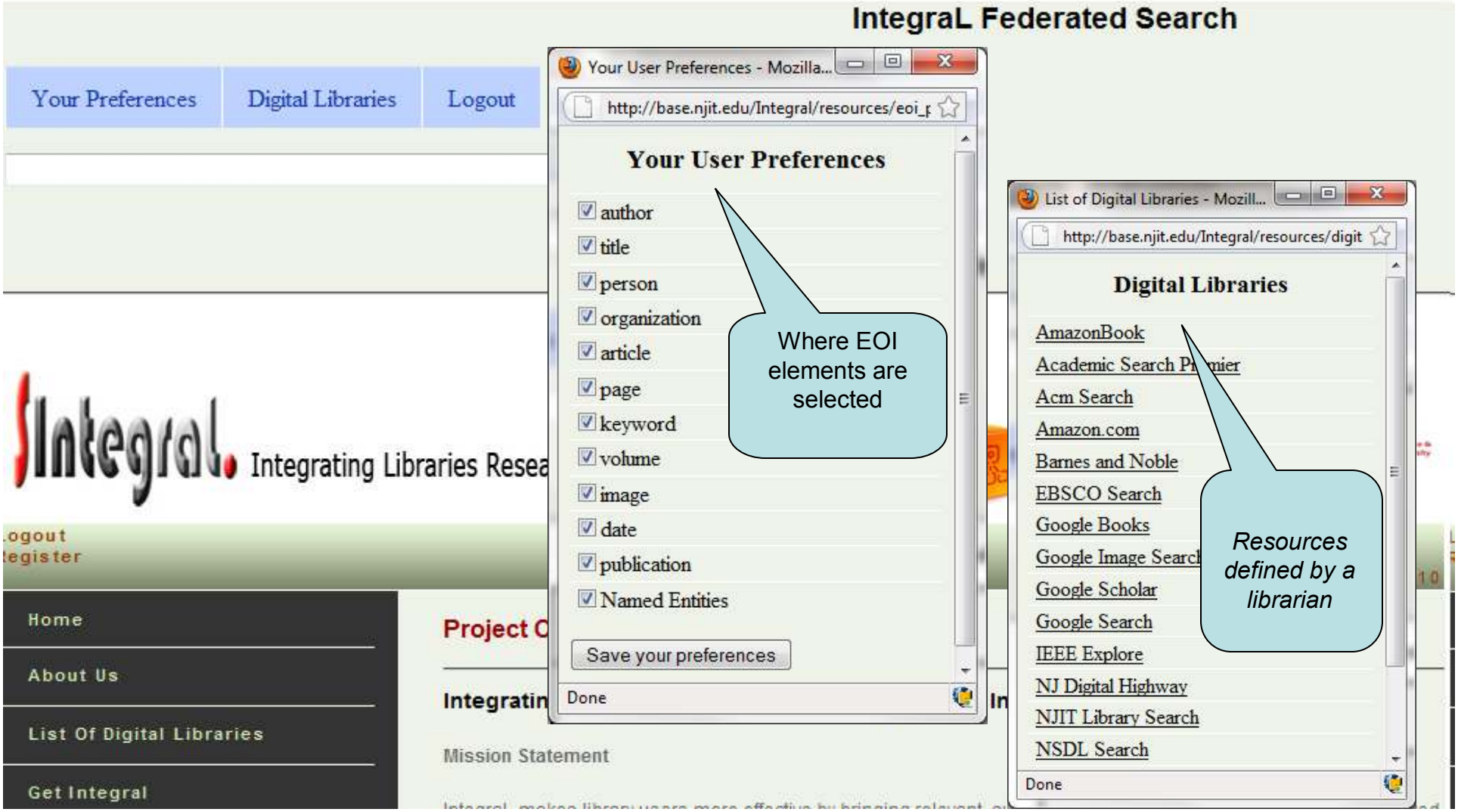
IntegraL Demo

NEWJERSEY Library Association

NJIT New Jersey's Science & Technology University

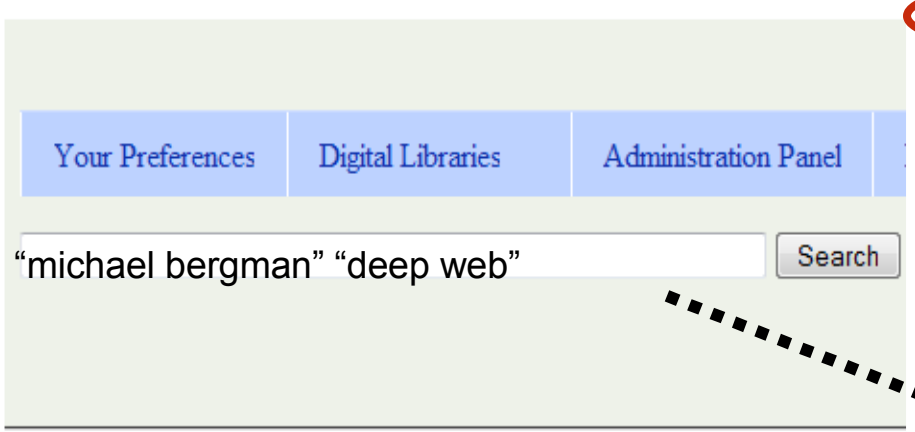
THE EDGE IN KNOWLEDGE

IntegraL Demo …



Integral Demo …

NEWJERSEY Library Association

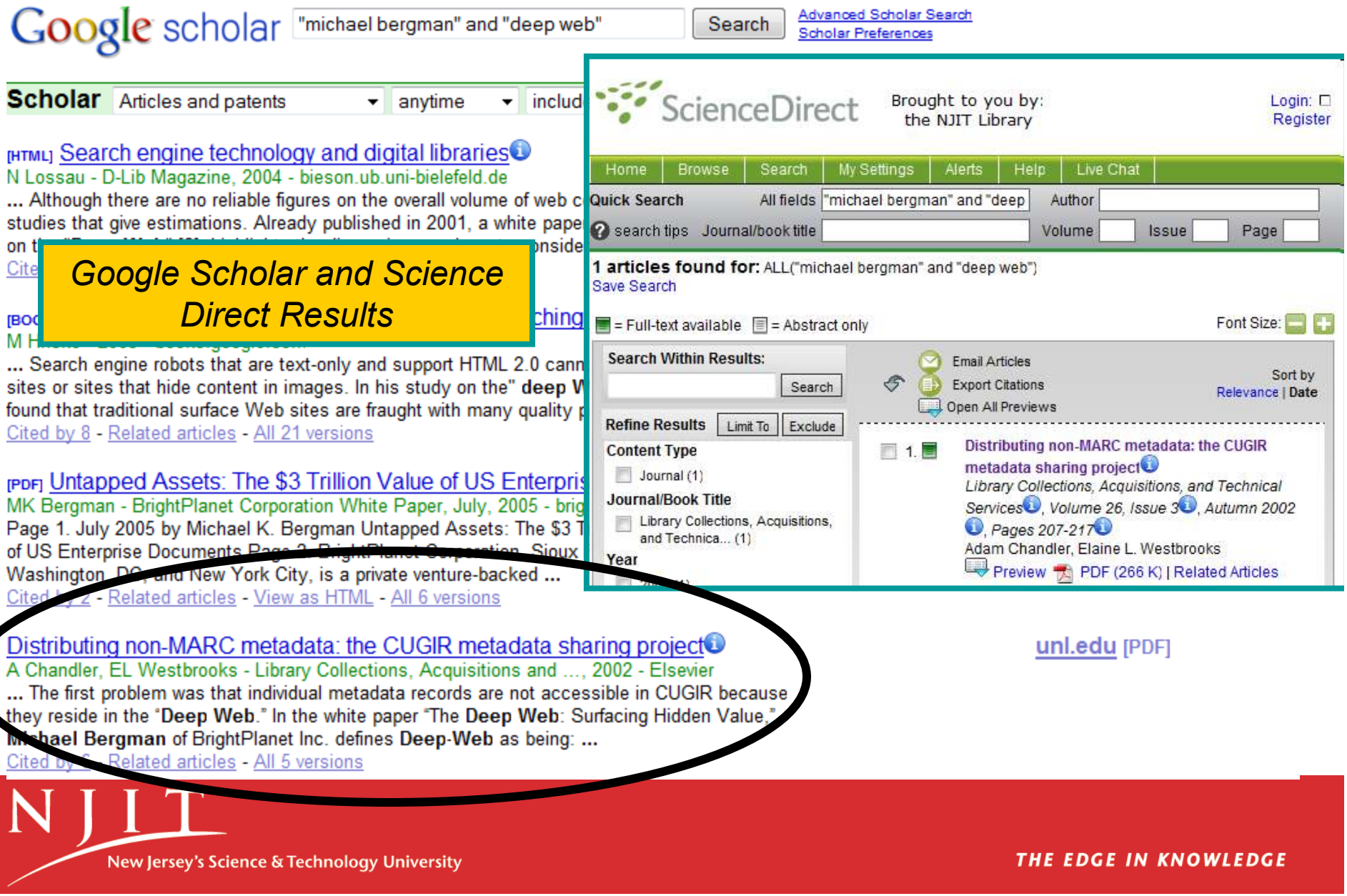


Federated Search Results

Google	4200 results
Ask Search	132 results
NJIT Library Author Search	0 results
NSDL Search	0 results
ACM Search	0 results
Google Book Search	0 results
GoogleImages	93 results
Scopus Affiliation Search	0 results
Google Scholar	40 results
IEEE Book Search	0 results
ScienceDirect	1 results

IntegraL Demo …

NEWJERSEY Library Association



IntegraL Demo …

NEWJERSEY Library Association

ScienceDirect

Brought to you by:
the NJIT Library

Home | Browse | Search | My Settings | Alerts | Help | Live Chat

Quick Search | All fields: "michael bergman" and "deep | Author | search tips | Journal/book title | Volume

Library Collections, Acquisitions, and Technical Services
Volume 26, Issue 3, Autumn 2002, Pages 207-217

Article | Figures/Tables | References | PDF (266 K)

doi:10.1016/S1464-9055(02)00247-6
Cite or Link Using DOI

Copyright © 2002 Elsevier Science Inc. All rights reserved.

Special section on metadata

Distributing non-MARC metadata: the CUGIR metada

Adam Chandler[a], and Elaine L. Westbrooks[b]

[a] Cornell University, Central Technical Services, 110c Olin Library, Ithaca, NY 1485

[b] Cornell University, Albert R. Mann Library, Technical Services, Ithaca, NY 14853-5

Available online 1 October 2002.

address two access-related problems. 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:

> … content that resides in searchable databases, the results from which can only be discovered by a direct query. Without the directed query, the database does not publish the result. When queried, deep Web sites post their results as dynamic Web pages in real-time. Though these dynamic pages have a unique URL address that allows them to be retrieved again later, they are not persistent [27].

Integral Demo

Google and Google Images Results

NEWJERSEY Library Association

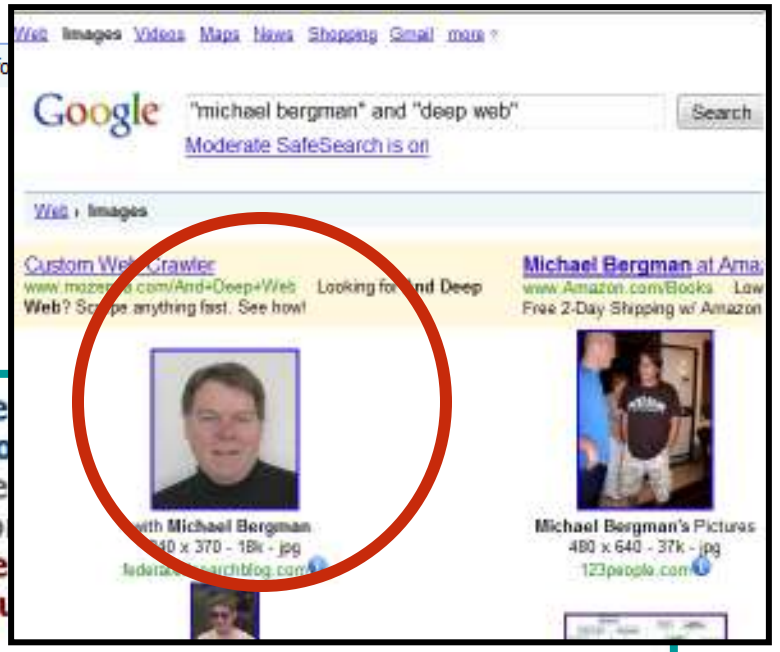
Google "michael bergman" and "deep web"

Web Show options... Results 1 - 10 of about 4,200 fo

AI3:::Adaptive Information » Mike Bergman on the semantic Web and ...
... Bioinformatics · Blogs and Blogging · Book Reviews · Brown Bag Lunch · **Deep Web** · Description Logics · Document Assets · Information Automation · irON ...
www.mkbergman.com/ - Cached - Similar

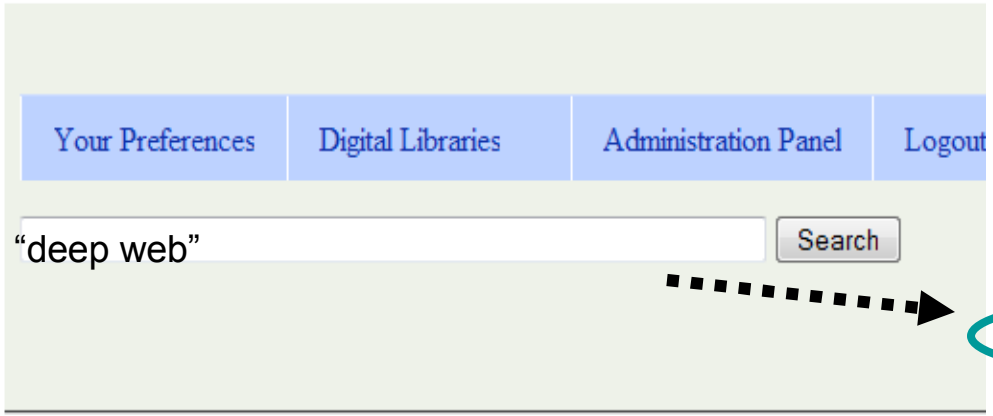
About Mike » AI3:::Adaptive Information
Nov 24, 2008 ... He was the author of BrightPlanet's award-winning Internet search tutorial and **deep Web** and document assets white papers. ...
www.mkbergman.com/about-mike/ - Cached - Similar

Cite
by M
Cite
be c
cites



Integral Demo …

NEWJERSEY Library Association



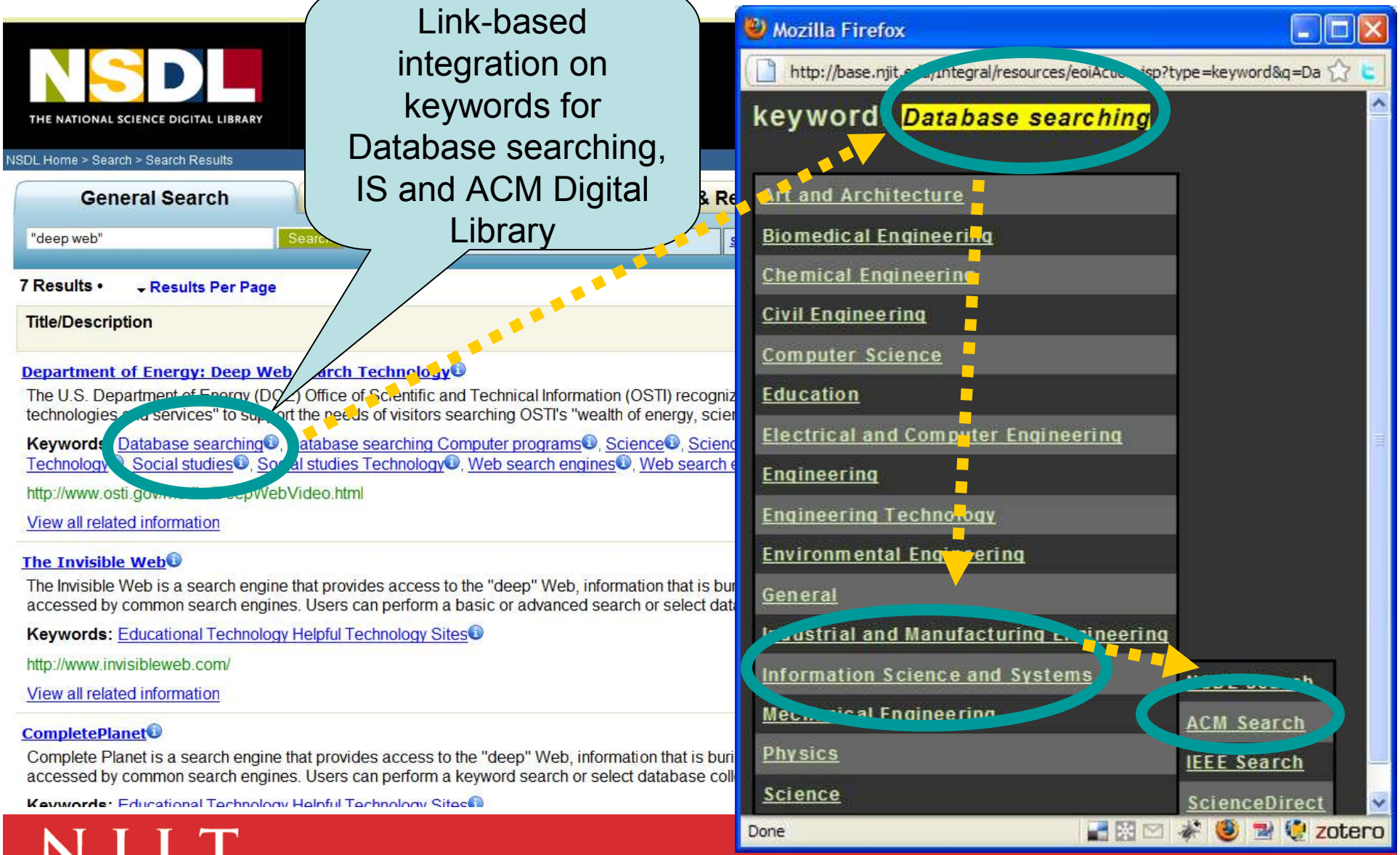
Federated Search Results

Google	245000 results
Ask Search	41900 results
NJIT Library Author Search	5599 results
NSDL Search	7 results
ACM Search	442 results
Google Book Search	0 results
GoogleImages	15100 results
Scopus Affiliation Search	0 results
Google Scholar	4280 results
IEEE Book Search	0 results
ScienceDirect	163 results

Integral Demo ...

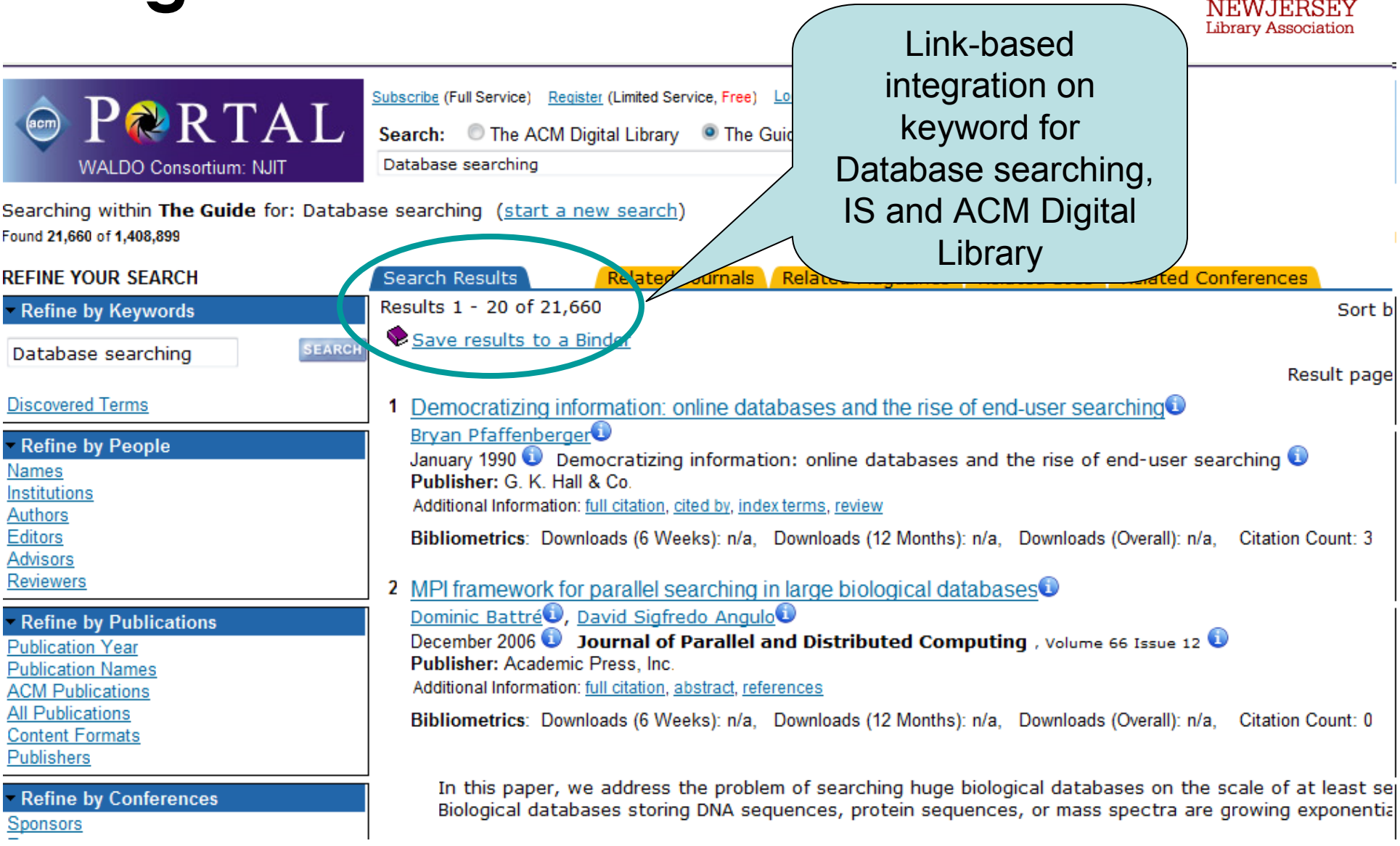
NEWJERSEY Library Association

Link-based integration on keywords for Database searching, IS and ACM Digital Library



IntegraL Demo ...

NEWJERSEY Library Association



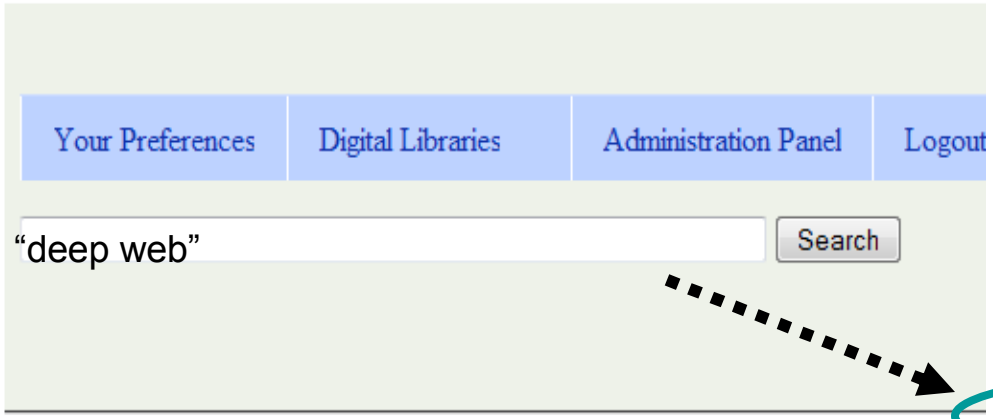
Link-based integration on keyword for Database searching, IS and ACM Digital Library

NJIT New Jersey's Science & Technology University

THE EDGE IN KNOWLEDGE

Integral Demo …

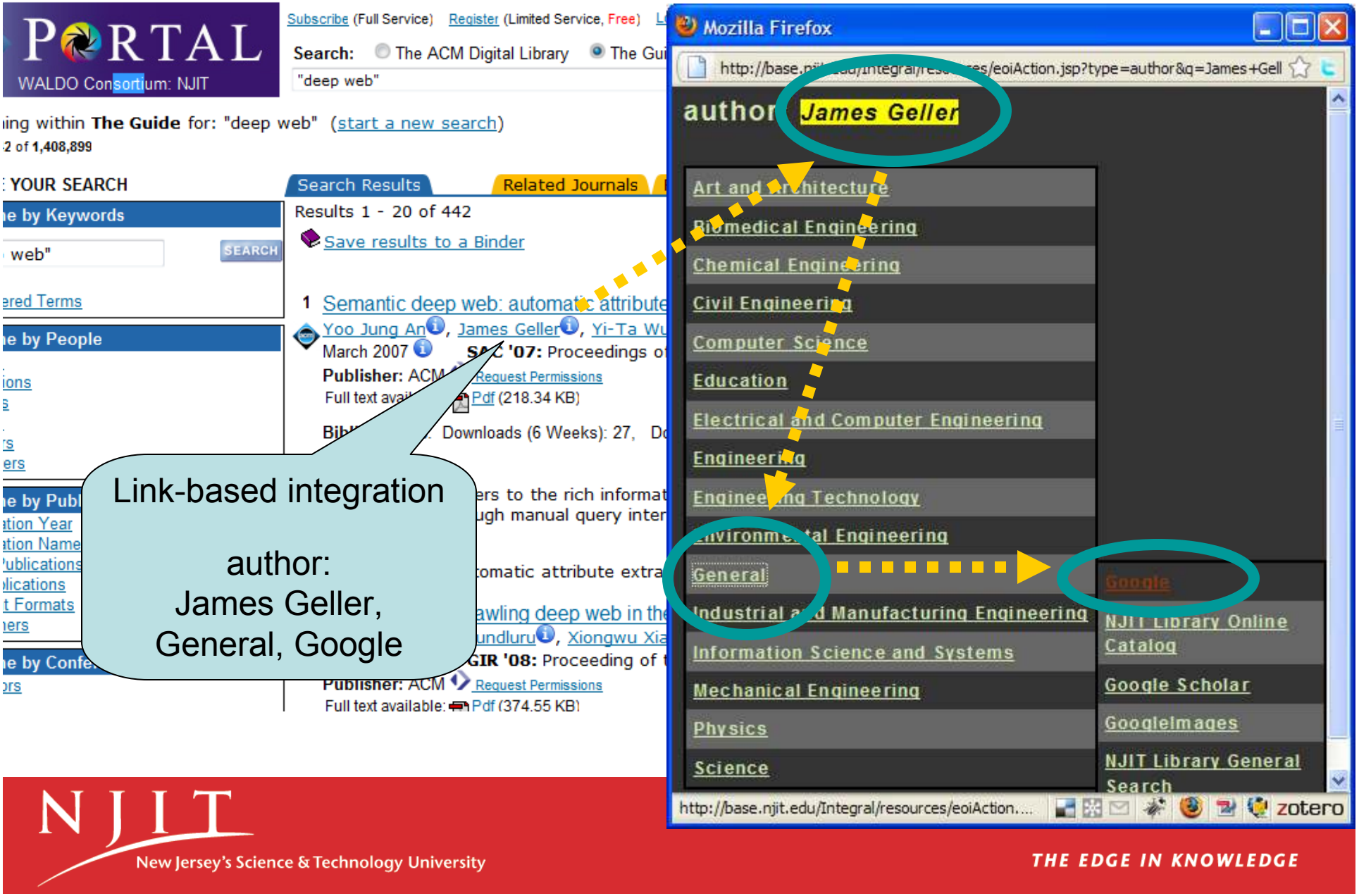
NEWJERSEY Library Association



Federated Search Results

Google	245000 results
Ask Search	41900 results
NJIT Library Author Search	5599 results
NSDL Search	7 results
ACM Search	442 results
Google Book Search	0 results
GoogleImages	15100 results
Scopus Affiliation Search	0 results
Google Scholar	4280 results
IEEE Book Search	0 results
ScienceDirect	163 results

Integral Demo ...



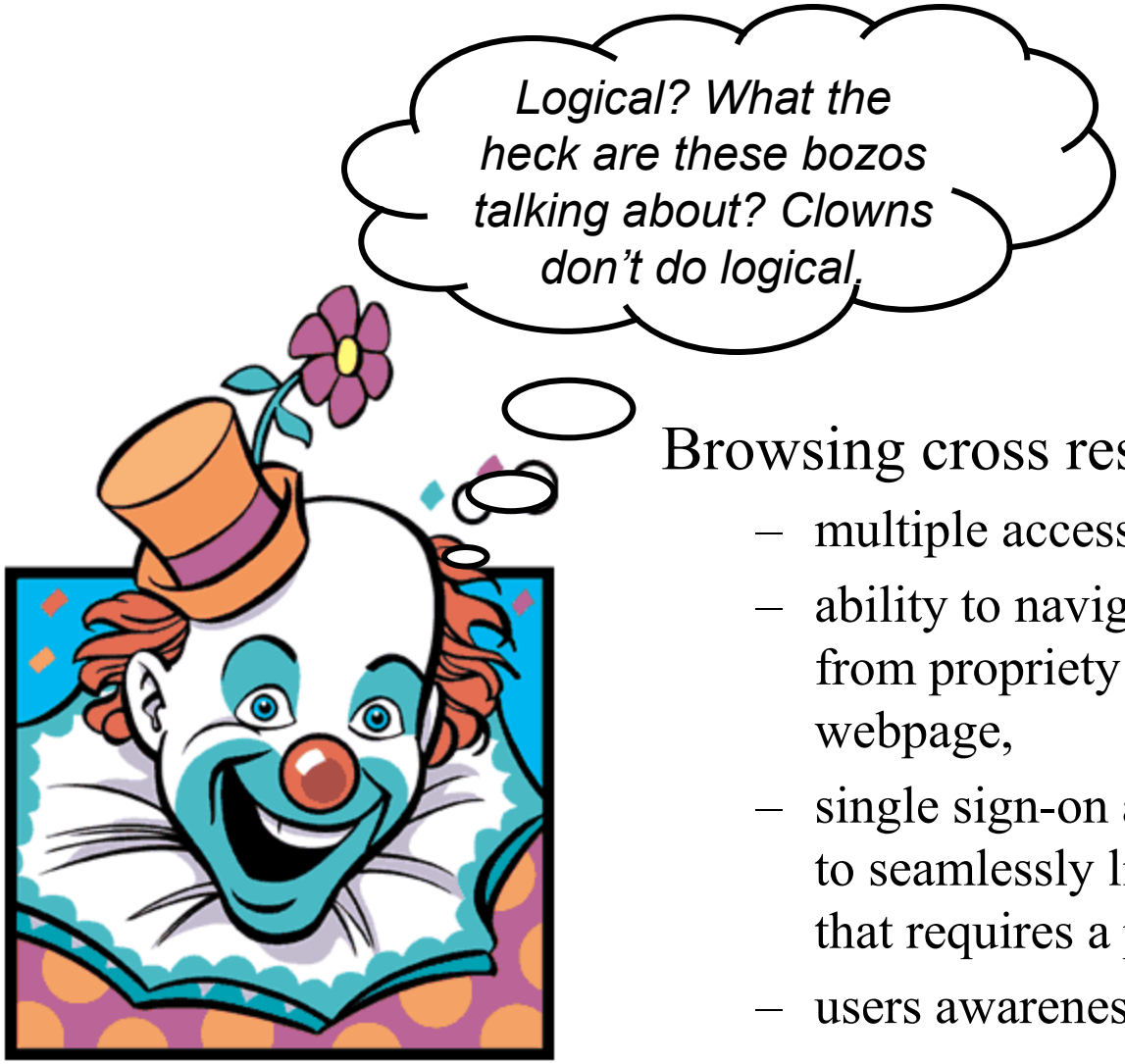
Integral Demo ...

NEWJERSEY Library Association

NJIT New Jersey's Science & Technology University

THE EDGE IN KNOWLEDGE

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

Benefits of Using IntegraL …

NEWJERSEY Library Association

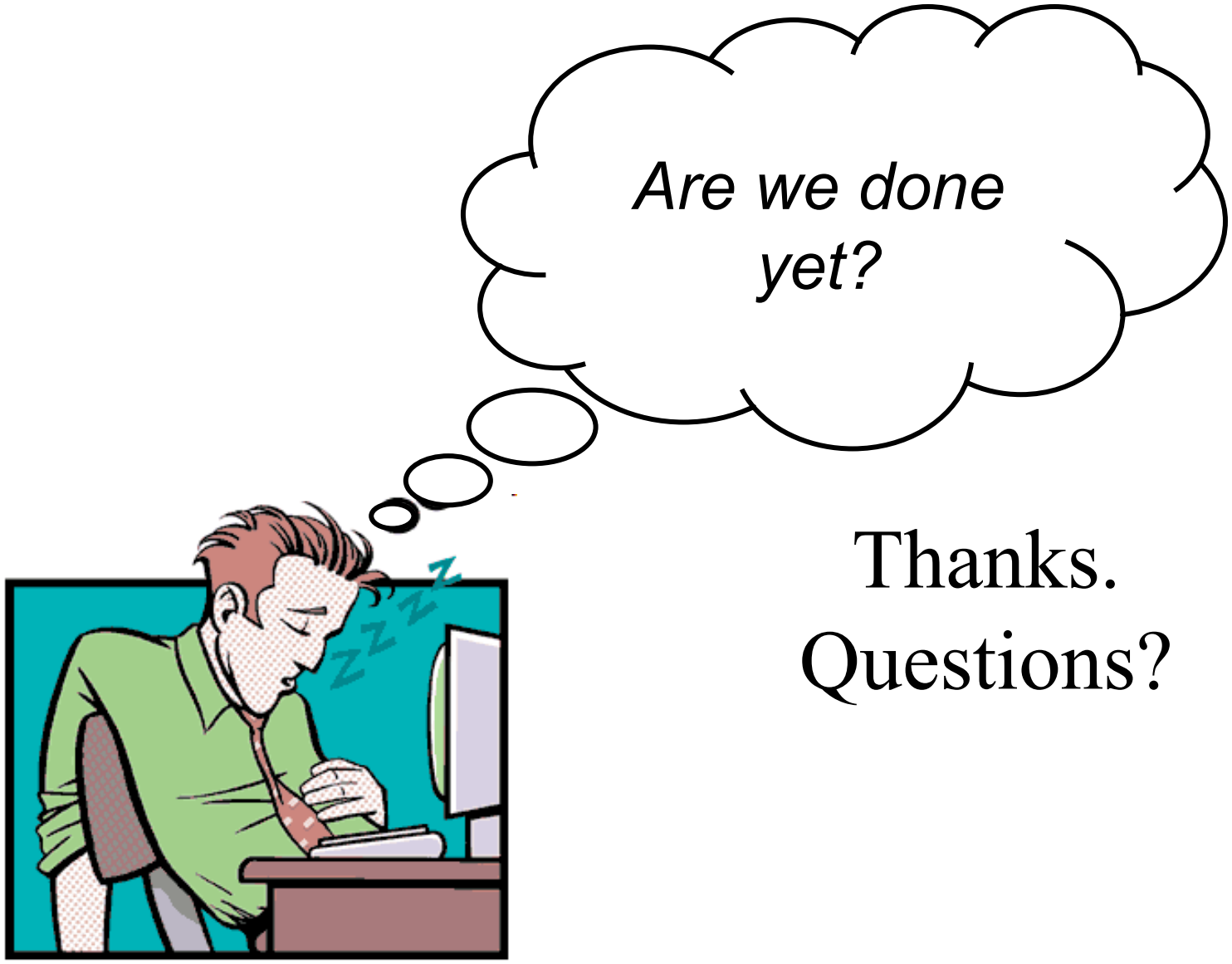
- Direct access to system's features
  - access to admin panel via the web
  - Ability add/modify/delete resources with minimal technical expertise.
- Information Literacy
  - IntegraL automates a search strategy that experienced searchers might go through by manually and inexperienced searchers might 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

NEWJERSEY Library Association

- Wikipedia Definition of "Deep Web" http://en.wikipedia.org/wiki/Deep_web
- Bergman, M. (2001). White Paper: The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, V 7(1). http://dx.doi.org/10.3998/3336451.0007.104
- Wright, A. (2009). Exploring a "Deep Web" that Google Can't Grasp. *New York Times*. Feb 23, 2009. *http://www.nytimes.com/2009/02/23/technology/internet/23search.html*
- 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
- Song, M. and Bieber, M. (2008). IntegraL: Lightweight Link-based Integration of Heterogeneous Digital Library Collections and Services in the Deep Web. E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008 10th IEEE Conference on Jul 21-24, 2008. http://doi.ieeecomputersociety.org/10.1109/CECandEEE.2008.111
- Song, M., Ho, S.M., Bieber, M., Koppel, E., Hamidullah, V., and Bokota, P. (2009). A Scalable Digital Library Infrastructure Expands Search and Beyond (2009) BookOnline'09 Workshop at ECDL 2009, October 2, Corfu, Greece

NEWJERSEY
Library Association



Thanks.

Questions?