Issues and Opportunities Associated with Federated Searching

Grace Baysinger (graceb@stanford.edu)
Head Librarian & Bibliographer
Swain Chemistry & Chemical Engineering Library
Stanford University Libraries and Academic Information Resources (SULAIR)

Abstract

Library catalogs and databases contain a wealth of information that is not available to Internet search engines such as Google. It can be difficult for users to identify which research tools to use and time-consuming for them to search each resource one at a time. Federated search tools make it possible to search multiple resources with one query. Several strategies have been developed to provide "one-stop shopping" but those dealing with multiple search interfaces face the biggest challenges. This talk will describe a project underway to develop federated searching prototypes on campus and will cover the viability of providing federated search services as well as the interest level by students and faculty in using them.
In May of 2007, SULAIR launched a *skunkworks* effort to push the envelope of the Libraries’ information discovery tools, services and environments. This effort is known as RaPIDS, for Rapid Prototyping of Intuitive Discovery at Stanford.

The primary objectives of the RaPIDS effort are to:

- Illustrate the features and functions of a cyber library, and
- Increase access to library resources.

This will be accomplished by:

- Rapidly deploying prototypes and gathering feedback, generating demonstrable results and tangible benefits, quickly,
- Focusing on delivering systems that meet patrons’ needs, and
- Helping catalyze the development of new library services.
RaPIDS: Team Members

- The effort will be led by representatives from each of the major units of SULAI R: Collections & Services, Technical Services and Academic Computing/Digital Library Systems & Services.

- Grace Baysinger, Head Librarian & Bibliographer, Swain Chemistry & Chemical Engineering Library
- Tom Cramer, Associate Director, Digital Library Systems and Services
- Adan Griego, Curator for Latin American & Chicana/o Studies
- Philip Scheur, Head, Cataloging and Metadata Services

- This core team will enlist colleagues throughout SULAI R for their expertise, ideas, content, technical know-how, and reviews.
RaPIDS Initial Projects

- Engaged in a sustained process of experimentation and progress, rather than the implementation of any single product.

- Examples of initiatives planned and underway:
  - Federated search
  - Associative searching
  - Taxonomic browsing

- RaPIDS page on Federated Search
  [http://library.stanford.edu/rapids/fedsearch.html](http://library.stanford.edu/rapids/fedsearch.html)
Federated Search

- Ability to simultaneously search multiple resources in real time and pool results in one merged relevancy ranked list.

- Gives scholars a broad view of disparate resources held across many different, isolated systems.

- SULAIR is collaborating with Deep Web Technologies. DWT’s Explorit Research Accelerator federated search engine is used in a number of science, technology and government search portals.
Other Portals Using Explorit
Why Use Federated Tools?

- Users don’t need to have prior knowledge about resources
- Help cope with lack of standardization in interfaces and search protocols for similar tasks
- Eliminates need to repeat same search multiple times
- Reduces time needed to find information
- Way to help users discovery high quality resources acquired by the library and to increase use of them
Three Demonstrations of Federated Search within Stanford Environment

Combined Search Engines Prototypes

Because Google is Just the Tip of the Iceberg, Explore the Depths

Discover Treasures in the Stacks

Encounter the Unexpected

“Top 10” Databases

All Library Catalogs at Stanford

Locally Digitized Collections

“Top 10” Databases

http://deepweb.stanford.edu/su/

All Library Catalogs at Stanford University

http://deepweb.stanford.edu/catalogs/search.html

Locally Digitized Collections

http://deepweb.stanford.edu/digcolls/search.html

Photo Credit:

chrisinh’s photos – Iceberg http://www.flickr.com/photos/chrisinh/page3/
terrytheschulz’s photostream – Diver http://www.flickr.com/photos/8745299@N03/342812239/
docman – Jelly-fish http://www.flickr.com/photos/docman/492317467/
“Top 10” Databases

RaPIDS Federated Search Prototype - Top 10 Databases

- ABI/Inform Global
- Annual Reviews
- BIOSIS
- Dissertations & Theses - A & I
- Engineering Village
- Expanded Academic ASAP
- Lexis Nexis Academic (News)
- Periodicals Archive Online
- PsychINFO
- Web of Science
All Library Catalogs at Stanford

Combined Search of All Library Catalogs at Stanford

Keyword: 
Title: 
Author: 
Match: All Field(s) 
Date Range: From Pick year To Pick year

Select All
- Socrates (Stanford's Online Catalog)
- Lane Medical Library Catalog
- Jackson Business Library Catalog
- SLAC Research Library Catalog
- Health Library Catalog (Stanford Hospital and Clinics)
Combined Search of Stanford Digitized Content

- Keyword:
- Title:
- Author:
- Match: All Field(s)
- Date Range: From To Pick year

Select All
- California Legislative Journal Appendices
- Copyright Renewal Database
- Corporation of Foreign Bondholders
- Dime Novels Full Text
- Documents to the People
- GATT Digital Library

- Joint Committee on Atomic Energy
- Medieval and Modern Thought Text Digitization Project
- R. Buckminster Fuller Collection
- Stanford University Publications
- SULAIR Books in the Public Domain
- Survey of Race Relations

© Stanford University, Stanford, CA 94305 (650) 723-2300. Terms of Use | Copyright Complaints
Rapid Retrieval, View Some Results While Search is Running
Sort Results – Rank, Source, Date, Title, Author
Ability to Select A Source
Help “Top 10” Databases

Combined Search of “Top 10” Databases-Bots - Description of Databases

https://idealsearch.stanford.edu/asz

**ABE/Inform Global (1993–)**

**Annual Reviews (1932–)**
Annual Reviews publishes authoritative, incisive reviews in 33 critical disciplines within the biomedical, physical, and social sciences. Annual survey publications are among the most highly cited in scientific literature. Since 1983 as of Oct 24, 2007, this is updated annually; articles for some issues are published online as soon as they are copy edited and peer reviewed.

**Biases (1986–)**
Biases is the major English-language service providing comprehensive worldwide coverage of research in the biological and biomedical sciences. It includes accounts of original research from nearly 3,000 primary journals plus meeting abstracts, reviews, books, book chapters, notes, letters, and technical reports. U.S. patents are included for 1952-1986, 1987-1990, and 1991 to 1994 in print. Size: 18 million records as of Oct 2006. Updated Weekly.

**Dissertations & Theses (Abstracts Only) 1861– (Online Full-Text 1997–)**
Includes citations to virtually every American dissertation accepted as an accredited institution since 1861. Selected Masters Theses have been included since 1975 and those from foreign countries since 1984. Also includes full text access for theabstracts only 1990 to present. Size: 5.01 million records as of Feb 2006. Updated: Monthly.

**Engineering Village (1884–)**
The most important resource in Engineering Village is Computer, the online version of Engineering Index. Computer is the most comprehensive bibliographic database of engineering research available today, covering more than 5,000 engineering journals, conference and technical reports. Size: 30.7 million records as of Feb 2007. Updated: Weekly.

**Expanded Academic ASAP (1983–)**
Balanced coverage is available through a variety of sources, including social science journals, humanities journals, science and technology journals, national news periodicals, general interest periodicals, and the New York Times (abridged only). Contains more than 1,500 indexed and full-text titles at which 2,100 are peer reviewed. Size: 111,691,050 articles as of March 2009. Updated daily.

**Lexis Nexis Academic (Data Covered Varies)**
This service provides full-text documents from over 5,000 news, business, legal, medical, and reference publications with a variety of flexible search options. Resources include: National and regional newspapers, wire services, broadcast transcripts, international news, and non-English language sources, U.S. patents and trademarks, case law, federal, state, and international case law, legal news, law reviews, and international legal information, Congressional Record, U.S. Supreme Court cases back to 1789, and Business news journals, company financial information, SEC filings and reports, and industry and market data.

**Periodicals Archive Online (1869–1995)**
An online archive of digitized, full-page journal articles in the humanities and social sciences, this resource provides access to international scholarly literature in the humanities and social sciences disciplines from 1869 to 1995. Currently contains 1.5 million articles from over 510 journals with links to JSTOR, Project MUSE articles. About 20% of the journals are in non-English languages. Size: 1.9 million articles as of March 2008. Updated quarterly.

**PsycINFO (1987–)**
The PsycINFO database provides access to the international literature in psychology and related behavioral and social sciences, including psychology, sociology, anthropology, education, pharmacology, and linguistics. PsycINFO contains citations and abstracts for journal articles, books, book chapters, reports, and dissertations. Journal journal segments substantive articles selected on the basis of relevance to psychology from over 1,700 journals published throughout the world in over 25 languages. Size: 243 million records as of July 2007. Updated Weekly.

**Web of Science (Science 1900–, Social Science 1957–, Arts & Humanities 1977–)**
The Web of Science is the premier and comprehensive source of social science, social sciences, and arts and humanities information from nearly 11,000 of the most prestigious, highly-cited research journals in the world. Web of Science also provides a unique search marketed, citation reference searching. A cited reference search enables you to find articles that cite a previously published work. In addition to citation reference searching, you can search three databases by topic, author, source title, and address. Size: Over 50 million records as of Oct 2006. Updated: Weekly.
Database Descriptions + Tags for “Top 10” Databases

Federated Search Prototype for the “Top 10” Databases

Database Descriptions Plus Subject Tags

**ABI/Inform Global (1923-)**
ABI/Inform provides broad coverage of business and management, including global business, company histories, competitive intelligence, and new product development. Over 2,600 publications, including full-text for 1,700 titles plus full-text for The Wall Street Journal, Eastern Edition; EU VentureWire; Geisinger Global Career Guides; Arthur Preface; and 14,000 ProQuest business dissertations. Size: 2.8 million records as of Jan 2006. Updated: Daily.

**Tags:** Accounting, Banking, Computer, Economics, Engineering, Management, Communications, Finance, Health Care, Human Resources, Insurance, International Trade, Law, Management/Management Science, Marketing, Public Administration, Real Estate, Tourism, Telecommunications, Transportation.

**Annual Reviews (1955-)**
Annual Reviews publishes authoritative, analytic reviews in 13 focused disciplines within the Biomedical, Physical, and Social Sciences. Annual Reviews publications are among the most highly cited in scientific literature. Size: 37,578 as of October 15, 2007. Site is updated monthly; articles for some series are published online as soon as they are copy-edited and typeset.


**BioMed (1964-)**

**Tags:** Aerospace Biology, Agriculture, Anatomy, Anthropology, Behavioral Sciences, Biochemistry, Biomechanics, Biotechnology, Biology, Cell Biology, Clinical Medicine, Environmental Biology, Experimental Medicine, Genetics, Immunology, Microbiology, Neuroscience, Pharmacology, Physiology, Public Health, Radiation Biology, Systematic Biology, Toxicology, Veterinary Science, Virology.

**Dissertations & Theses (Abstracts & Indexes 1861 - ; Stanford Full-Text 1989-)**
Includes citations to virtually every American dissertation accepted at an accredited institution since 1861. Selected Master’s Theses have been included since 1962 and those from foreign sources since 1985. Also includes full-text access for Stanford theses from 1989 to present. Size: 2.02 million records as of Jan 2006. Updated: Monthly.

**Tags:** Agriculture, Astronomy, Biological and Environmental Sciences, Business and Economics, Chemistry, Education, Engineering, Fine Arts and Music, Geography and Regional Planning, Geology, Health Sciences, History and Political Science, Language and Literature, Library and Information Science, Mathematics and Statistics, Philosophy and Religion, Physics, Psychology and Sociology.
Summary Description & Features

- Single search to heterogeneous set of resources
- Federated search is a discovery tool, not for comprehensive searching

- Results in real time from resources, not snapshots
- Fast retrieval and display of some results while search is still being completed

- One ranked list after merging results
  - Up to 100 records per resource
  - Ability to sort several ways, including source

- Link to full record from resource
  - “Permanent Link” is cache of full display

- Change preferences after doing an initial search
Technical Challenges in Using Federated Search Tools

- Speed and performance of other federated search engines
- Rapidly growing number of digital resources
- Type of searchable information in resources varies
- Resources w/ limited numbers of simultaneous users or require individual login before use
- Nearly constant change in interfaces, some changes in URLs
- Names of search fields and formatting of data in them are not standardized
- Sorting order varies for default display of results
- Number of records that can be displayed w/o violating license agreement or slowing search engine too much
- Need to retain “state of session”
- Clearing cache between sessions
Preliminary Feedback

- Almost all students really liked using federated search prototypes and found results to be relevant.

- Librarians want more precision for author searches, made suggestions for improving the interface. Much discussion of choice of “top 10” files. Mixed value of results from library catalogs prototype.

- Strategically, federated search is one of the directions that we want to pursue.
Future Plans

- Offer more resources via federated search
  - eBook & other full-text resources
  - Abstract & Indexing Databases
  - Special Collections Finding Aids
  - Reference resources & web sites

- Implement new interface offered by DWT
  - More search fields
  - Faceted search results
  - Exporting tagged results
  - SFX linking
  - Alerts
Acknowledgements

- Stanford: RaPIDS Members - Tom Cramer, Phil Schreur, Adan Griego; other staff who helped implement the prototypes; plus the many staff and students at Stanford who tested and provided feedback.

- Deep Web Technologies: Abe Lederman, Sol Lederman, Brian Desplain, and other members of the DWT team.