Metasearching Lessons
from the California Digital Library

Roy Tennant
California Digital Library
Metasearching at CDL: A Brief History

- 1997, UC San Diego creates *Database Advisor*
- 1999, CDL decides to create a metasearch tool based on DBA, to be called SearchLight
- 2000, SearchLight is launched
Welcome to SearchLight - Sciences/Engineering

With SearchLight, you can search many of the databases and other resources available to CDL users - all at the same time. It can find books, journal and encyclopedia articles, and quality Internet sites. SearchLight will run the search and bring back the results. With SearchLight, you don't have to worry about finding the best databases first - it does it for you. For social sciences and humanities topics, go to SearchLight - Social Sciences/Humanities.

SearchLight will search ALL science/engineering databases in the CDL profile. If you wish, you can target your search to a more specific list of databases by selecting the specific subtopics at the right: [Explain]

Enter your search terms:

for example: HIV protease inhibitor

How many minutes are you willing to wait? 1 min [Explain]

http://searchlight.cdlib.org/cgi-bin/searchlight
**Sciences/Engineering**

You searched for: tamoxifen

Refine your search? | Help | Feedback

---

**UC San Francisco**

Remember – not everything is available online. There are also a number of print resources in the UCSF Library that can lead you to useful information. See important print resources at UCSF.

---

**Jump to:**

**Books**
Library catalogs and other book citations

**Full-Text Resources**
Electronic texts and documents

**Journal Indexes**
Indexes and abstracts to journal articles

**Reference Resources**
Encyclopedias, dictionaries, directories

**Electronic Journals**
Full-text online journals

**Web Directories**
Links to Internet resources

---

**Books**

Jump to: Books | Journal Indexes | Electronic Journals | Full-Text Resources | Reference Resources | Web Directories

---

**# Hits** | **Resource Name & Description**
---

**35** | **WorldCat** [more info]
Citations for books, journals and materials in other formats

**16** | **MELVYL Catalog** [more info]
Library catalog for all UC campuses
### Journal Indexes

<table>
<thead>
<tr>
<th># Hits</th>
<th>Resource Name &amp; Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8767</td>
<td>BIOSIS Previews [more info]</td>
</tr>
<tr>
<td></td>
<td>Life sciences and biology; indexes over 6000 journals, plus books and conferences</td>
</tr>
<tr>
<td>6787</td>
<td>MEDLINE/HealthSTAR [more info]</td>
</tr>
<tr>
<td></td>
<td>Citations to journal articles in all areas of medicine and health sciences</td>
</tr>
<tr>
<td>2229</td>
<td>Current Contents [more info]</td>
</tr>
<tr>
<td></td>
<td>Covers 6500 scholarly journals in all disciplines from 1989+</td>
</tr>
<tr>
<td>278</td>
<td>MAGS: Magazine &amp; Journal Articles [more info]</td>
</tr>
<tr>
<td></td>
<td>Citations to general-interest articles, many with abstracts, some with full text</td>
</tr>
</tbody>
</table>

### Electronic Journals

<table>
<thead>
<tr>
<th># Hits</th>
<th>Resource Name &amp; Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>678</td>
<td>Elsevier ScienceDirect [more info]</td>
</tr>
<tr>
<td></td>
<td>Electronic journal articles published by Elsevier</td>
</tr>
<tr>
<td>450</td>
<td>Academic Press (IDEALIBRARY) [more info]</td>
</tr>
<tr>
<td></td>
<td>Electronic journal articles published by Academic Press</td>
</tr>
<tr>
<td>324</td>
<td>Kluwer Online [more info]</td>
</tr>
<tr>
<td></td>
<td>Electronic journal articles published by Kluwer Academic Publishers</td>
</tr>
<tr>
<td>45</td>
<td>JSTOR [more info]</td>
</tr>
<tr>
<td></td>
<td>Electronic journal articles from many publishers provided by JSTOR</td>
</tr>
</tbody>
</table>
Metasearching at CDL: A Brief History

- 2001-3, CDL looks beyond the limitations of the SearchLight service
- 2002-3, Commercial products become viable
- 2003, CDL Metasearch Infrastructure Project formally launched
Lessons from SearchLight

- **Metasearching is worth doing** (it’s what many users want)
- For a large research library, **metasearching is best focused on particular needs or subject areas** (the “drinking from a firehose” problem)
- **Not all databases are created equal** (e.g., we need a way to focus on core databases, or have results from core databases ranked higher than others)
- Focusing on specific audiences or needs provides an opportunity to expand service beyond simple searching
<table>
<thead>
<tr>
<th>Select</th>
<th>Database</th>
<th>Summary</th>
<th>Coverage</th>
<th>Full Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Academic Search Elite</td>
<td>General information from journals, magazines, and newspapers in all subject areas.</td>
<td>1984 - present; fulltext, 1990 - present.</td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>AccuNet/AP Photo Archive</td>
<td>News photos from the 1840's to today are covered. This resource is updated continuously.</td>
<td>1840's - present</td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>American National Biography Online</td>
<td>Profiles of 18,000+ men and women who have influenced American history and culture.</td>
<td>Colonial-present</td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>ArticleFirst</td>
<td>Business, Humanities, Medicine, Popular Culture, Science, Social Science, Technology.</td>
<td>1990 - present</td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>Biography Resource Center</td>
<td>Worldwide coverage of past and present notable people. Narrative biographies, key facts, selected articles, and more.</td>
<td></td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>Ethnic NewsWatch</td>
<td>The ethnic, minority, and native press in America.</td>
<td>1985 - present</td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>Funk and Wagnall's New World Encyclopedia</td>
<td>General Encyclopedia.</td>
<td></td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>Informe (English)</td>
<td>Hispanic reference material covering cultural and lifestyle issues, health and parenting, entertainment, politics and current events.</td>
<td>1996 - present</td>
<td>Partial</td>
</tr>
<tr>
<td></td>
<td>JSTOR - The Scholarly Journal Archive</td>
<td>Electronic access to older scholarly journals.</td>
<td>Varies</td>
<td>Partial</td>
</tr>
</tbody>
</table>
CIAO Focus, January 2004: India and Pakistan

This month's meeting between the leaders of India and Pakistan has renewed hopes that a real and lasting peace can be made between the two nuclear powers. Prime Minister Atal Behari Vajpayee and Pakistani President Pervez Musharraf put forth a new initiative aimed at normalization of relations. In a sentiment echoed by Musharraf, Vajpayee called for a bold transition. "We have to change South Asia's image and its standing in the world," Vajpayee said. The dispute over Kashmir, which has sparked two of their three previous wars, remains a major stumbling block. Separatist violence continues to be on the rise in Kashmir, with over a dozen people killed in December despite a truce. The recent rapprochement has been greeted with optimism by the world community.

From CIAO's database:

Finding a Kashmir Settlement: The Burden of Leadership

The Start of a Beautiful Friendship? The United States and India

Is South Asian Nuclear Politics Different?

Some Cooperative Initiatives for India-Pakistan Dialogue and Risk Reduction

Why India Went Nuclear

Politics in Pakistan Post-September 11, 2001

Outside Links:

United Nations Military Observer Group in India and Pakistan
Principles

- Metasearching is a tool that is appropriate for some tasks, but not for others
- Only librarians like to search, everyone else likes to *find*
- All things being equal, one place to search is better than two or more
- “Good enough” is often just that
Principles

- The size of the result set isn’t as important as how the results are displayed (the Google lesson)
- Our ability to create effective one-stop searching is dependent on our ability to appropriately target user needs
- Services should be placed as close to the user as possible
Building a Metasearch Infrastructure

- Goal: to build a reference implementation that allows us to “eat our own dog food” while providing:
  - An implementation for campuses that don’t want to mess with it
  - An easy-to-modify set of templates for those who do
- Requirements: shared resource descriptions, templates, implementation guidelines, etc.
The Metasearch Infrastructure Project

The Metasearch Infrastructure Project aims to create an infrastructure that can be used by campus libraries and CDL to tailor search portals to particular user needs. Metasearching has also been called "cross-database searching" and "federated searching". This infrastructure would replace the existing SearchLight service with a "family" of portals aimed at particular audiences and needs. Based on a thorough review of the marketplace, we are presently working to acquire a commercial product by first quarter 2004. Contact Roy Tennant for more information.

Our vision of how this tool could be deployed within the university information environment is depicted in the diagram to the right (click on it for a larger version). The "Scholar's Box" refers to a tool being developed by the U.S. Bureau of Education Information Systems. Our metasearching infrastructure as a resource for university faculty, K-12 teachers, and others to use to create learning objects, exhibitions, and other types of products based on selected and

www.cdlib.org/inside/projects/metasearch/
## EZArticleFinder

Search for **giant squid** found 60 results. 1-50 of 60 records retrieved.

<table>
<thead>
<tr>
<th>No.</th>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th>Source</th>
<th>View abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Miller GA</td>
<td>The magical number seven plus or minus two: some limits on our capacity for processing information.</td>
<td>1956</td>
<td>Psychol Rev [via Medline Plus]</td>
<td>View abstract</td>
</tr>
</tbody>
</table>
Database Advisor Service

User

Initiates search

Sends search to

Metasearch Software

Sends search to

Database Advisor Tool

Performs search, identifies top 2-3 DBs, writes out file referenced by results page

Launches multiple searches

Receive results

Merges, deduplicates results

Builds display

Sends merge display to

Sends search to

Interface software

Send search to

Database Advisor Service

Send search to

Database Advisor Tool

Performs search, identifies top 2-3 DBs, writes out file referenced by results page

Launches multiple searches

Receive results

Merges, deduplicates results

Builds display

Sends merge display to

Sends search to

Interface software

Send search to

Database Advisor Tool

Performs search, identifies top 2-3 DBs, writes out file referenced by results page

Launches multiple searches

Receive results

Merges, deduplicates results

Builds display

Sends merge display to

Sends search to

Interface software

Send search to

Database Advisor Tool

Performs search, identifies top 2-3 DBs, writes out file referenced by results page

Launches multiple searches

Receive results

Merges, deduplicates results

Builds display

Sends merge display to
Plenty of Problems to Go Around

- Database Vendor Issues
- Software Provider Issues
- Library Issues
- User Issues
Database Provider Issues (the short list)

- Access control (robust authentication and authorization)
- Load
- Inappropriate searches (searching databases that don’t apply)
- Branding and “unfair” deduping
Software Provider Issues (the short list)

- Access management
- Search mapping
- Unreliability of targets
- Systems that don’t support an API (that must be screen-scraped)
- Inadequate result data for good:
  - Deduping
  - Ranking
Library Issues (the short list)

- Selecting the right system
- Cost (both upfront and ongoing)
- System design and implementation
- System maintenance
  - Ability to add new resources/targets
  - Ease of interface changes and upgrades
- System inadequacies (e.g., ranking)
User Issues (the short list)

- What must I go through before hitting the search button?
- How difficult is it to review results?
- Are results ranked by relevance? (that will be my assumption)
- Will I get buried? (too many sources, too many results?)
- Do I have methods to easily focus in on what I want?
- Once I find what I want, can I get to the full-text with a click?
- Can I copy a citation and put it in my paper?
Challenges

- Software still needs improvement (understatement)
- Some databases are still not searchable
- How do we create an infrastructure that is easy to deploy for a variety of purposes and/or audiences?
- We can learn from other systems (e.g., redlightgreen)?
- Standards are on the way (e.g., NISO), but many of us have deployed systems that could use solutions now
Metasearching Today & Tomorrow

Today:
- One-stop shopping
- Broad subject categories with many databases each
- Not integrated with any other system (e.g., course management systems)

Tomorrow:
- Tailored portals for specific user needs or topic areas
- Targets created by libraries for specific purposes (e.g., focused crawling of web sites, harvesting of repositories)
- Dynamic selection of sources based on user