Alan Manifold
Systems Implementation Manager
Purdue University Libraries

ACRL Delaware Valley Chapter
19 November 2004

Breaking the Google Addiction
the promise and reality of federated searching
Who am I?

Just a Techie, ENCompass only

Not!
OUTLINE

- Google, et al
  Pros & cons, new stuff, failure rate
- Promise/Vision of Federated Searching
  Definitions, screen shots
- Reality of Federated Searching
  Problems, successes, challenges
- Some Reflections
What does Google provide that libraries have not traditionally provided?

**GOOGLE – PROS**

**Ranking – keeps track of popularity**
... the order of records returned is based on the number of links to it on the web ...

**Relevancy – analyzes the content**
... the order of records is also based on their relevance, determined from the content ...

**Really easy to use**
... just type in a word, Google does the rest ...

**Comprehensive**
... it hardly matters what topic or area you want, Google has scads of hits for it ...

**Consistent in response**
... Google basically always returns the same kind of screen ...

**Nice looking**
... it’s colorful, it’s got a cute name, it has indentations and headings, etc ...
Google’s biggest failings:

Selection: No selection criteria to insure quality, accuracy, peer review, etc.

Authority control: No control over content, which can be inaccurate and inconsistent, as well as being filled with words added just to affect relevancy score

Currency: Returns countless dead links

Weighting: Can weight HTML title tags more than other content, but can’t do much beyond that

Cataloging: No descriptive metadata to describe document content overall

GOOGLE – CONS
Still listed more than two years after going 404

Dead links seem like a serious problem, but nobody seems to care, right? What can we learn from that?

GOOGLE’S FAILURES
Google is best at fact finding at a shallow depth:
Who starred in that?
When did that happen?
What’s its atomic number?
Where can I get one?
Who makes those?
What are its side effects?
How do you do that?
When is it showing?

Google is not so good at depth and analysis:
Tell me about it.
Why did they?
How are they related?
Which is better?
What do the experts think?
Has that been proven?
How do we know that?
What led up to that?
We worry when our children date someone we think is superficial and shallow.

We worry about our patrons when they rely totally on Google for their information needs when it does not provide the depth and quality of information they could be getting.
Initially, federated searching was simply the ability to search a number of disparate resources with a single search.

Current federated searching products often also include support for:

- Multiple protocols
- Multiple data formats
- Multiple search types
- Results consolidation
- Record de-duping
- Results sorting
In the future, federated searching products could expand into powerful searching tools.

Already they are evolving towards the ability to support such things as:

- Searching and presenting video and audio files
- Personally contoured searching
- Searching numerical data
- Searching and presenting non-textual data (e.g. maps, genomes, chemical compounds)
- Institutional repositories
These features rely on searching object metadata. As the ability evolves to search object attributes directly, and peripheral options expand for presentation, federated searching systems should be there:

- Search for and present aromatherapy solutions
- Search webcams by image attributes
- Search and reproduce holographs
- Search (and even clone) genomes
- Search and reproduce chemical compounds
- Search parts and assemblies by shape

Distant Future?
Federated Searching is a more powerful tool than Google in some significant ways:

**Federated Searching**
- Dynamic
- Multiple protocols
- Open
- Focused

**Google**
- Static
- HTTP protocol
- Proprietary
- Unfocused

**IS FS THE ANSWER?**

-Breaking the Google Addiction
- the promise and reality of federated searching
The Internet is based primarily on HTML, which codes information for display:

```
<table border="0" cellpadding="2" cellspacing="3" class="fixed" width="100%">
<tr><td valign="top" class="dl">Database</td><td valign="top" class="dt">Academic Search Elite</td></tr>
<tr><td valign="top" class="dl">Title</td><td valign="top" class="dt">Male red-sided garter snakes, Thamnophis sirtalis parietalis, determine female mating status from pheromone trails.</td></tr>
<tr><td valign="top" class="dl">Creator</td><td valign="top" class="dt">O'Donnell, Ryan P.<br> Ford, Neil B.<br> Shine, Richard</td></tr>
<tr><td valign="top" class="dl">Source</td><td valign="top" class="dt">Animal Behaviour &nbsp; &nbsp;Oct2004, Vol. 68 Issue 4, p677</td></tr>
<tr><td valign="top" class="dl">Notes</td><td valign="top" class="dt">...snip...
[Copyright 2004 Elsevier]</td></tr>
</table>
```
Records in most library systems are coded in MARC, which works well for describing traditional library material content, but is difficult to extend to other material types:

OCLC MARC Bib Record in Raw Form:
00734cam 22002411 45*0001001300000003000600013005001700019008004100036
0100017000770400023000940430012001170500016001290820013001450920019001
5804900090017710000250018624500980021126000570030930000390036635000090
0405504003000414651004800444*ocm00442080 *OCoLC*19940620065418.0*7010
12s1968 pauab b 000 0 eng * ‡a 68021623 * ‡aDLC‡cDLC‡dOCL‡dIPL‡fan-us
---*0 ‡aJK2556‡b.E2* ‡a325.3/73* ‡a325.373‡bEb61f* ‡aIPL1*1‡aEblen, Jack Eri
cson.*14‡aThe first and second United States empires;‡bgovernors and territorial gover
nment, 1784-1912.*‡a[Pittsburgh]‡bUniversity of Pittsburgh Press‡c[1968]* ‡aviii, 34
4 p.‡billus., map‡c24 cm.* ‡a8.95* ‡aBibliography: p. 321-333.* 0‡aUnited States‡x
Territories and possessions.*
Federated searching products use XML, which codes information for content, not for display.

```xml
<MARC>
<MRleader>02412naa 2200289 4500</MRleader><MR001>14582694</MR001>
<MR008>200410e20041001xxu####e###j########eng#d</MR008>
<MR022><MR022a>0003-3472</MR022a></MR022>
<MR072><MR072a>Article</MR072a></MR072>
<MR100 ind1="1" ind2="0"><MR100a>O'Donnell, Ryan P.</MR100a></MR100>
<MR700 ind1="1" ind2="0"><MR700a>Ford, Neil B.</MR700a></MR700>
<MR700 ind1="1" ind2="0"><MR700a>Shine, Richard</MR700a></MR700>
<MR245 ind1="1" ind2="0"><MR245a>Male red-sided snip</MR245a></MR245>
<MR270><MR270a>Dept of Zoology, Oregon State U</MR270a></MR270>
<MR514><MR514a>Peer Reviewed</MR514a></MR514>
<MR520><MR520a>...snip...</MR520a></MR520>
<MR654><MR654a>GARTER snakes</MR654a></MR654>
<MR903><MR903a>20041001</MR903a></MR903>
<MR945><MR945m>68</MR945m><MR945n>4</MR945n><MR945p>677</MR945p></MR945>
</MARC>
```
EnCompass federated search by title screen

Select search options

Select databases to search

SAMPLE SCREEN #1
EnComPass federated search by subject screen

- **Expand subjects**: Click to expand the subjects list.
- **Select DBs**: Select the databases you want to search.
- **Arts and Humanities**
- **Business and Economics**
- **Engineering and Technology**
- **General and News**
- **Government and Law**
- **Life Sciences**
- **Medicine and Health**
- **Physical Sciences**
- **Social Sciences and Education**

---

**SAMPLE SCREEN #2**

*Breaking the Google Addiction: the promise and reality of federated searching*
Federated search results in progress

SAMPLE SCREEN #3

Breaking the Google Addiction: the promise and reality of federated searching
1. Results limited for better response time
2. We don’t know which 34 hits and we can’t get the ones after 100 (the limit)
3. It got to the database, but the search itself failed
4. It didn’t connect to the database
5. Click to get usually useless messages to explain failures (e.g. “Unknown error”)

NEEDS EXPLAINING
<table>
<thead>
<tr>
<th>Source</th>
<th>Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspaper Source</td>
<td>Democratic presidential field on health care MICHELE NORRIS; All Things Considered (NPR); 07/07/2003</td>
<td>07/07/2003</td>
</tr>
<tr>
<td>Newspaper Source</td>
<td>Democratic hopefuls clash over war, health care Steven Thomma; Knight Ridder Tribune Washington Bureau (DC); 06/03/2003</td>
<td>06/03/2003</td>
</tr>
<tr>
<td>Newspaper Source</td>
<td>Democrats on Stump Plot Their War Rhetoric Glen Johnson; Boston Globe, The (MA); 03/11/2003</td>
<td>03/11/2003</td>
</tr>
<tr>
<td>Academic Search Elite</td>
<td>An analysis of the perceptions and characteristics of... Dinnebell, Laurie A.; Topics in Early Childhood Special Education; Summer98, Vol. 18 Issue 2, p118</td>
<td>1998</td>
</tr>
</tbody>
</table>

“Show All” results screen – no sorting specified
The record display screen shows selected fields from record, as determined by the library. The presence of the button means there is a URL in the record.
Although federated searching has a number of successes to its credit, there are also a number of problems. And since the products are relatively immature, there are many more challenges ahead. Let’s look at:

😊 Successes
😊 Challenges
😊 Problems

REALITY
A number of problems contribute to the difficulty of making federated searching match its vision. Among them:

- Lack of standards
- Multiple protocol support
- Multiple data formats
- Range of vendor support
- Search definitions
- Z39.50 problems
- HTTP Search Engine (HSE) connectors
What does a title search cover…
On my online catalog?
On another online catalog?
On an A&I database from vendor P?
On an A&I database from vendor E?
On Amazon or Barnes & Noble?
On Associations Unlimited?
On Google?
On Funding Opportunities Database?
On Reference Suite@FACTS.com?
On Biography Resource Center?
With these Z39.50 Attributes: Use:4, Relation:3, Position:3, Structure:1, Completeness:1, Truncation:1?
HSE Connectors have as their mission to extract specific data fields whether they are there or not.
HSE Connectors do their work by emulating a web browser. They connect to web pages, read the HTML and try to interpret what they read. And some programmer has to tell them how.
Database vendors complain about HSE connectors because they pound the web servers much more than individual users could do.

System vendors hate using HSE connectors, but the database vendors have not provided alternatives, such as XML gateways or Z-connections.

V Said:  
Z Said:
CHALLENGES

Authorization
Make sure only authorized users can get to specific resources

Connectors
Keep current connectors working and move away from HSE connectors to something more stable

Response time
Search multiple resources faster

Integrating new resources
As new protocols and resources come into being, federated search systems need to keep up

De-duping and managing results
When results are like apples and oranges, sorting and de-duping are tough, but the users expect it
IT REALLY WORKS!

At Purdue, we have 119 databases listed on our “MegaSearch” pages, about half HSE and half Z39.50

Endeavor claims 138 ENCompass sites
Ex Libris claims 531 MetaLib sites
MuseGlobal MuseSearch (couldn’t tell)
Sirsi SingleSearch (couldn’t tell)
WebFeat claims 1500 sites

SUCCESSES
To attract our patrons to use federated searching, we need to address Google’s strengths head-on.

**IS THIS ENOUGH?**
Maybe we should also look at the problem differently:

We try to get people to come to the library, but maybe a better model for the web would be to put what the library offers into places people are already going. What would this look like?
Both Google and Yahoo now include links to WorldCat. Type “find in a library” with any search to get library info:
The bottom line is value. We need to give our patrons valuable services in exchange for their time and effort.

Federated Searching adds value to the library’s offerings. Patrons get more results per minute spent.
Picture a patron doing federated searching and getting one-click access to:

- Full text of journal articles
- Photos, graphs, maps, video clips, sounds
- E-Books full text
- Data sets in an institutional repository
- Selected peer-reviewed websites
- Dictionaries and encyclopedias
- Specialized software

**ENVISION**
When we deliver that, our patrons will rush to choose federated searching over Google.