eResearch Australia

Open Access Archiving and other Gateways to e-Research

Richard Levy

The Elephant in the Room!
The Impact of Google on eResearch

- “Google is the ‘black box’ to information on the Internet, providing simultaneous searching of millions of resources in a convenient user friendly way. A ‘one-stop-shop’ for information retrieval. “ (Miller)

- “The quality and content of information retrieved may be in question, as may be the authenticity and credibility of resources. Nevertheless it leaves end users with information, choices and possibilities that they can then process for themselves, and utilise or not utilise as the case may be.” (Craven)

The Impact of Google Scholar

- “Google Scholar software can analyze all of this harvested data and metadata at its leisure and store the results in advance of actual searches, instead of needing to perform on-the-fly processing while a user waits for results. This preprocessing can be applied to all the articles in the local index, not just the first 50 articles to be returned by each source in a cross-search” (Rothkind)

- “The most obvious and frequently cited problem with Google and Google Scholar is the lack of exhaustivity: you can't know precisely what the Google index includes and what it leaves out. There's no guarantee that all of your library's expensive licensed content, which you want to make sure users can find, is included”
Find Research with Google

Google Scholar

Advanced Scholar Search

Subject Areas:
- Biology, Life Sciences, and Environmental Science
- Business, Administration, Finance, and Economics
- Chemistry and Materials Science
- Engineering, Computer Science, and Mathematics
- Medicine, Pharmacology, and Veterinary Science
- Physics, Astronomy, and Planetary Science
- Social Sciences, Arts, and Humanities

Google Scholar
Google Scholar

E-Resource Access and Management Services

Can Google be improved upon?

I first met Buckminster Fuller at my apartment about three years ago. He and his wife then lived in the same building I do. An ISI colleague, Beta Starchild, told me that Fuller needed a new office. The University of Pennsylvania could no longer provide space for located in the University City Science Center.

Fuller was born in Milton, Massachusetts, on July 12, 1895, the second of four children. His mother was Caroline Wolcott Fuller, and his father Richard Buckminster Fuller, Sr. The Fuller

E-Resource Access and Management Services
Web 2.0

- The Web has changed how we deliver and consume information
  - The shift from physical to digital delivery of information has created new requirements and opportunities for delivering effective library experiences

- The Web has profoundly transformed the nature of library collections
  - The majority of new acquisitions are Web-based
  - Collections have increased dramatically and content is available anytime, anywhere
  - Web search engines compete with libraries

Open Archive Access (OAI-PMH)

- Harvesting metadata from multiple sources
- Typical of the institutional repository (Open Source and Hosted)
- Essential to Google and Google Scholar but…
- Also key to other harvesters such as OAIster, Thomson Citation Index, SCOPUS
- Creating local indexes, storing metadata
- Still limited by publishers and licensing issues, hence Open URL
- Bridging OAI and Open URL is still in its infancy, e.g. via CrossRef DOI
The Differences

- Two traditionally separate areas of content management
  - Institutional Repositories provide free open access articles and objects published by the institution
  - Link Resolvers provide access to content usually only available via journal subscription
  - Repositories use XML to publish content to the web
  - Link Resolvers based on Open URL or DOI
  - Content from Repositories is harvested (OAI)
  - Content from link resolvers is usually restricted to patrons and is not retrievable unless freely available to the web
**The Differences**

**IR**
- Open Access
- Free
- Harvested
- Stored in one place
- Owned by institution: no copyright restrictions

**LS**
- Open URL
- Subscription
- Restricted access
- Published by multiple providers
- Not normally owned by institution: copyright restriction

**University of Hokkaido Airway Project: 2**

1. OpenURL Request
2. OpenURL Request
3. XML
4. 1CATE's link navigation window created

Knowledge Base

CrossRef

1CATE (link resolver)

1CATE's
(link resolver)

Metada accumulated preliminary

HUSCAP
Another IR
Another IR
Another IR
Is Federated Searching the Answer?

- Federated searching aggregates multiple channels of information into a single point.
- It can search metadata and full-text.
- It can partner with a link resolver to access complete articles.
- It can utilise clustering to categorise results.
- It can search proprietary, local and open sources (Google).
- It incorporates A&I databases, OPACs, repositories.
- Not just a search engine but a unified search interface:
  - Xml gateways
  - Z39.50
  - HTTP
  - Knowledge Base

Searching without Federated Search

Searching Individual Resources – One at a Time

Given a keyword or topic, patrons don’t know where to begin amongst all the available resources.

- Time consuming and inefficient.
- Many interfaces to learn – quality varies.
- Richness of collection is lost.
The Limits of Federated Searching

- “Meta-searching…is a step backward, a way of avoiding the learning process” (Slaney)

- “You can't get better results with a federated search engine than you can with the native database search. The same content is being searched, and a federated engine does not enhance the native database's search interface. All federated search does is translate a search into something the native database's engine can understand.” (Hane)
Perceived Limits of Federated Search

- Relatively limited syntax
- Unseen translator algorithms
- Restricted access to subject
- No thesauri or controlled vocabulary
- Limited number of results per search
- Not always compatible with native database indexing
- No subheadings or secondary filters
- “Creates Finders as opposed to Searchers”

Fed Search versus Single Database

- Example: looking for the adverse effects of Prozac
- Native databases via aggregator like OVID, DataStar
  - Embase
  - Medline
- Syntax
  - Concept 1. – generic drug name
  - Concept 2. – adverse effects
- Use of online thesaurus or thesaurus mapping
- Run equivalent search in federated search resource
- Run search in Google Scholar
Outcomes

- **The native database** provides for a more focused and subject-specific result with filtering to thesaurus terms (generic drug name, drug subheading)
- **Federated searching** results relevant but more limited in number and variable: can only approximate the single database level of precision
- Clustering mapped drug name to some medical/pharmaceutical descriptors* but searching on the generic name for drug improved result
- **Google Scholar** results relevant but deteriorated towards bottom of list suggesting typical Google phrase-indexing methodology
- *in Medline & Ovid
Search Example 2

Looking for articles on “Chromatography”

Syntax
- Use the Cross-Search facility in the traditional database aggregator to locate relevant sources
- Run a comparable search using federated searching resource
- Run the search in Google Scholar
Fed Search
E-Resource Access and Management Services

Outcomes

- The search via the traditional **database aggregator** did not give a clear indication of the type of results only the volume
- Was unable to search the entire collection: had to specify a subject area
- No links to abstracts, metadata, full text links*
- **Federated searching** used automatic clustering to indicate the topics, sources, authors and date of publication
- Sources were clearly visible and accessible
- Links were available to abstracts and full-text
- **Google** does not provide indexed results, advanced Boolean
- Google Scholar range of subjects limited
  
  * Using the Cross-Search facility
Conclusions

- Googlisation, for lack of a better word, is inevitable
- Helping users to find resources in one place is preferable if not essential
- A unified resource that does this is an advantage
- Fed Search is a way-in not a cop-out, a gateway and a place to gather relevant research that may require further/deeper investigation into specific databases or aggregated collections
- Like consulting a broad encyclopedia divided into several subjects versus a subject-specific textbook where the indexing is more specific
- More cross-referenced metadata is needed to enhance the potential of federated searching in the future
- This, however, applies to virtually all aggregated collections as each database is idiosyncratic and few are fully integrated

Conclusions

- Federated searching applies across disciplines and subjects and can serve a wider research community
- Many end-users and even information professionals find it difficult to keep up with each and every source of data
- A resource that provides a substantial cross-section of research from every available source adds value to the complete collection
- It does not necessarily replace the ongoing need for user education, induction and current awareness of what sources and strategies may be appropriate for research
- Federated searching is a hybrid approach, mimicking Google yet providing access to other sources, including subject-specific databases
- Less a one-stop shop and more of a shopping mall
What applies to Google…

- “…it leaves end users with information, choices and possibilities that they can then process for themselves, and utilise or not utilise as the case may be.”
  (Craven)

Thanks

richard.levy@anz.proquest.com