Lost in translation: another view on metasearching

Dennis Warren,
La Trobe University

Abstract

Metasearch systems promised to solve all of the problems presented by unique vendor driven search interfaces. But since the metasearch "solution" was proposed, the environment has changed considerably. This paper discusses the changes that have occurred (the enhancements offered by vendor platforms to take account of feedback from users, the provision of cross-searching via vendor platforms, and the emergence of Google Scholar) and proposes that the metasearch "solutions" that have been built to date have failed to provide a better alternative.

Introduction

What commenced as a simple idea – allow the end-user to perform a search simultaneously across a range of commercial databases – has, in many ways, proved difficult to implement. While it was easy enough to produce a single search box, it has proved difficult to produce meaningful results at the end of the process. There is no doubt that user expectations have been raised as a consequence of widespread and positive experience with Google, and more recently with Google Scholar. But, to date, it is not clear that the results from metasearching implementations are meeting the needs of these same end users.

Those libraries who have been willing to report on their implementation of metasearch applications have described missed deadlines, soft launches and compromises made along the way and much of the effort has gone into what one commentator has called the "plumbing layer", where time is spent building connections to diverse resources (Jenkins as cited in Rogers 2006) p. 27. In one of the most detailed descriptions to date of the implementation of a metasearching system, Highsmith and Ponsford have described the resistance of library staff to the resultant system. They have noted that many library staff had "higher pre-implementation expectations than the software could support" and the library staff "object(ed) to the way in which federated search systems "dumb down" native interfaces." (Highsmith and Ponsford 2006) p. 193.

This paper explores pre-implementation promises, and compares that with current experiences for the end user. It also compares the sophistication of native interfaces with the unsophisticated nature of searching in the metasearch environment. In the implementation process, many compromises have been made, and the question that has to be asked is whether the current implementation of metasearching is close enough to the original idea to be worth continued development effort. Or have we lost sight of the original idea? Is near enough, good enough? Or has the idea been "lost in translation"?

LibXplore

La Trobe University and eleven other universities are part of the Australian Academic and Research Library Network (AARLIN) using Metalib from Ex Libris and the OpenURL link resolver SFX as their portal solution. The early phases of the AARLIN project have been documented elsewhere (Gow and Roberts 2003), (Burke, Fisch et al. 2004). The number of university libraries participating in the project changed from 22 to 12 when the project commenced its current phase in 2006.

The local La Trobe University version is known as LibXplore and there are institutional level variations both in what the local implementation is called (Cross Search, The Library Portal, Supersearch etc) and how it is configured.

At the time of writing, there were 528 "active" resources configured for LibXplore. Table 1 gives the breakdown by type of resource.

Category of Resource	**Number of Resources Configured**
Database	293

E-Journals	35
Electronic Books	4
Encyclopaedias	4
E-Prints	1
Images	14
Library Catalogues	89
Newspapers	1
Search Engines	3
Subject Gateways	80
Theses	4
	528

Table 1: Resources configured for LibXplore

The examples used in this paper relate only to resources configured as “databases”. The terms “database” and “subject gateway” are used loosely in this environment. For example, subject gateways are the 7:30 Report (ABC), 60 Minutes (Channel 9), ABC Radio National and At the Movies (ABC).

There are two issues with regard to configuring a large number of resources for federated searching. The first issue is the time taken to configure each resource. The second issue is the issue of sustainability. Bigger is not necessarily better, and configuring a resource “just because you can” is hardly a persuasive argument for inclusion. Error messages are already not uncommon and there is no evidence that systematic procedures are in place to ensure that configuration issues are quickly resolved.

Promise and reality

The merits of metasearching, from a theoretical perspective, have been explored in a number of papers. Much has been promised but, to date, the reality has fallen short of that promised. It is possible to discern the following themes (promises) in the literature. These vision statements are then compared with the current implementation.

1. Federated searching will allow users to search a range of databases simultaneously. According to the proponents of this view, users “have been frustrated with the complex array of databases they see on most library web pages, and would rather have ‘one-stop shopping’ for their information needs” (Oberhelman 2006) p. 6.

   Reality: As implemented at La Trobe University, federated searching allows the end user to search up to 10 databases simultaneously. As Figure 1 indicates, of the active resources in LibXplore, only 62% of the resources configured are fully searchable and capable of presenting the results in the LibXplore environment. In other words, 38% of the resources identified as important cannot be fully integrated into the metasearch environment.

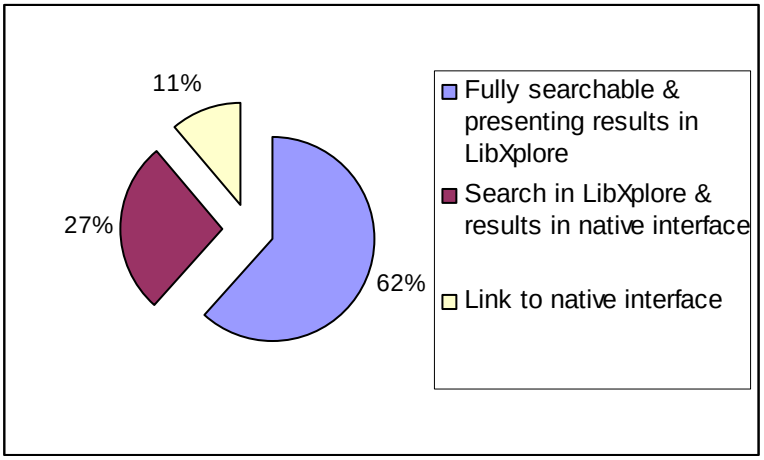

Figure 1: Resources by Configuration Type

While users can build their own set of (up to 10) databases to search, novice users are expected

to make use of “quicksets”, or pre built topical collections, of databases. Figure 2 illustrates a search, using the Law quickset, which is made up of five databases, for the phrase “card fraud”.

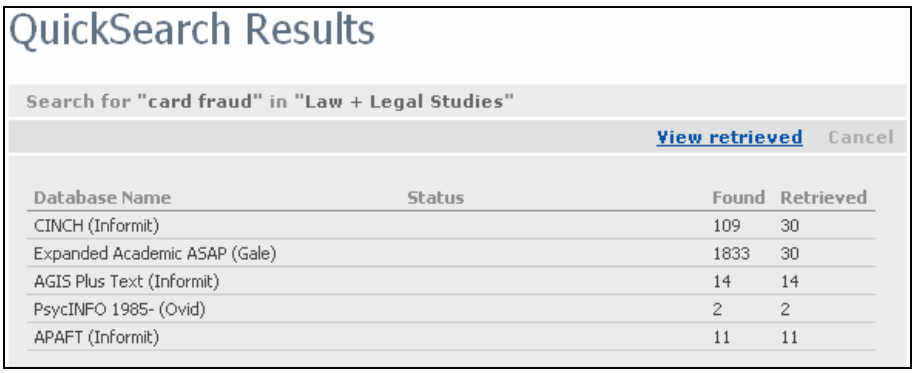

Figure 2: A search using the Law + Legal Studies quickset

As the above search illustrates, the quicksets are often made up of both full text databases (Expanded Academic ASAP) and indexes. When you include a full text collection in your set of databases to be searched the results from the indexes will be less numerous, and the ranking algorithm will ensure that the results from the full-text collection out rank all other results.

Presenting users with a pre built subject quickset runs the risk that end users will not make use of the most appropriate databases for a particular topic. For example, one of the best databases for students studying family law is Family, the database produced by the Australian Institute of Family Studies. For dispute resolution, you would have to include CRInfo, a free database that is tailored for the topic. For intellectual property law, one of the best databases is Sniper, and for banking and finance law, ProQuest has some merit. It is not possible to build a legal quickset that takes account of the best databases in all the areas of law (because of the limit of 10 databases) so metasearching doesn’t really solve the problem of getting the end user to the most relevant database(s). As a “standalone” resource discovery tool it is no better than the long alphabetical lists arranged by topics that we have previously offered. As noted above, there are 293 databases configured for LibXplore. How can we be sure that users get the right ten?

To emphasise this point further, does this environment make it easier for the art historian (or art history student) to find the art history databases? The answer is clearly no! If the end-user knows that art history falls in the domain of the humanities (and there would be some art history students who do not see this connection) and selects the Humanities quickset, the topic is run through the seven databases that make up the Humanities quickset (see Figure 3). But, is there any point running an art history search through Linguistics and Language Behavior Abstracts or Sociological Abstracts? Wouldn’t it be more relevant to include Bibliography of the History of Art in the quickset?

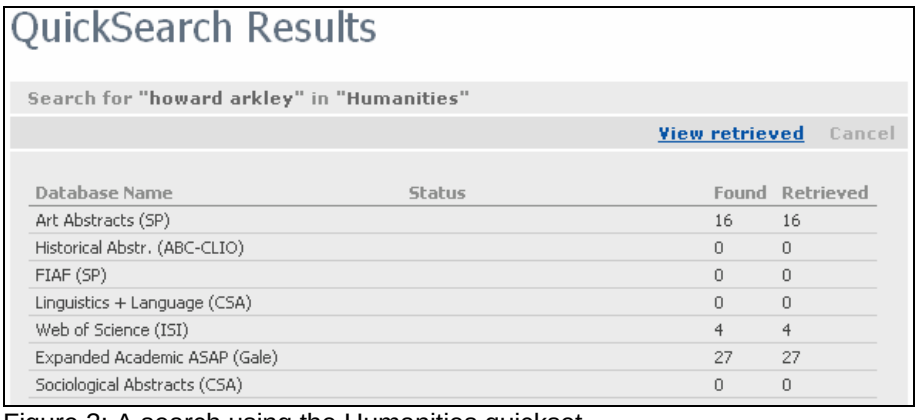

Figure 3: A search using the Humanities quickset

Is the option to build more quicksets? While that might be desirable it is hardly practical, in a large university teaching a broad range of courses. As Tallent has noted when reporting on a

MetaLib implementation at Boston College,

”(S)tudents were in as difficult a position as they were (using) the library’s list of online databases when it came to resource selection. If a student saw a database called Francis, it meant little in both environments, as students are generally unaware of the scope of resources available, do not recognize appropriate databases by title, and the interface does not make it easy to select an appropriate database….To date, the most noteworthy and popular enhancement….is the quick search functions.” (Tallent 2004) p. 74.

It is impossible to have quicksets for all law subtopics and all humanities disciplines yet this is precisely what the end-user requires.

It should be acknowledged that it is possible, using the Find Database functionality to navigate to an Art History list (see Figure 4) but using this approach a user is then presented with a list of 16 “resources” and no obvious way to form a sub-set to search. Also, there is no attempt to rank the resources – the Bibliography of the History of Art does not leap out from this list as a quality art resource, and there are no annotations/descriptions (other than the map of Australia if it is relevant to the content) on this top level page to identify the “best” database(s). It is possible to change the user preferences so a brief annotation for each database displays but, at the time of writing, each user is required to set their own display preferences individually.

One could well ask why we persist in presenting our users with alphabetical lists of databases, especially since many of the names of the databases tell us nothing about the content? This is clearly a step backwards. While the student can access a full description of each database, how many will read the descriptions of the 16 resources before making their selection(s)?

Database Name	Type	Actions
AHB (Informit)	Database	
America: History and Life	Database	
APAFT (Informit) Full Text	Database	
ARCH (Informit)	Database	
Art Abstracts (SP)	Database	
Berlin University of the Arts	Library Catalogues	
Bibliog. History of Art (RLG)	Database	
Bibliography of Asian Studies	Database	
Design + Applied Arts Index	Database	
Expanded Academic ASAP (Gale) Full Text	Database	
Google Image Search	Images	
Handbk Latin American Studies	Database	
Index Islamicus (CSA)	Database	
INFOMINE: Visual & Performing Arts	Subject Gateways	
Medieval Music Database	Database	
Proquest Social Science Full Text	Database	

Figure 4: Art History “resources”

As Frost has argued, improving the beginning researcher’s selection of research tools is part of the educational process. According to Frost,

”We already have many ways to assist database users. Libraries …need to develop web sites that are well organized…; web and/or printed handouts to guide searchers; knowledgeable and friendly reference staff available most library hours; comprehensive instruction programs; and good relationships with classroom teachers.” (Frost 2004) p. 68.

In the metasearching environment, this process is not made any easier. It can be argued that it has been, in fact, made more difficult.

2. Patrons have stopped using resources because they are frustrated by the number of dissimilar search interfaces they must use to access content. (Marshall, Herman et al. 2006) p. 173. "The job of information professionals is not to make all users into information professionals. Our job is either to give them the right tools for the job or do the job for them" (Miller 2005) p. 58. According to the proponents of this view, federated searching eliminates dissimilar search interfaces and metasearching will entice a generation of Google fans back to paid library databases. Putting this line of thought succinctly, Marshall et al have argued:

   "(F)ederated search provides a way for librarians to reclaim the community that Google snapped up" (Marshall, Herman et al. 2006) p. 180.

   Reality: Do we have evidence that end users have felt frustrated and have stopped using library databases? While a number of studies have indicated that search engines are the first port of call for many students (see Bawden and Vilar 2006 for a comprehensive review), there is also considerable evidence that usage of databases in libraries was going up, not down, prior to the introduction of metasearching. As we saw above, because not all databases can be configured for metasearching, the user will continue to have to master a range of interfaces. And the reality (explored below) is that the user interfaces have become much better. Indeed, by comparison with the metasearch interface, the native interfaces offer a level of sophistication that our users both want and can master. There is a risk that, rather than winning back the Google fans, clumsy metasearching implementations will drive frustrated users more quickly to Google (and Google Scholar) and away from paid library databases.

   To what extent does metasearching offer end users the Google experience? In many ways LibXplore has more in common with traditional database searching than with Google. Enter two words, and LibXplore searches for that phrase whereas Google does an automatic "and" search. Some databases now mimic Google and, increasingly, two search terms will be automatically "anded" without the need to enter the Boolean operator. A search for the words "veterans vietnam" in Historical Abstracts results in 14 records, but the same search in LibXplore results in 0 records, because LibXplore interprets two words as a phrase. Searching Historical Abstracts, via the native interface, includes automatic truncation and there appears to be a developing trend, driven by user experiences with Google, towards automatic truncation. Via LibXplore it is necessary to use the asterisk for truncation (but there is no onscreen help to tell users what the truncation symbol is). The Advanced search screen in the metasearch environment looks more like the native interface (without any bells and whistles), and less like the Google and Google Scholar approaches.

   There are many other ways in which Google and LibXplore are different. Google allows for spelling errors, and suggests preferred spelling, via the "Did you mean?" functionality. Google does not require the user to select resources to search. Google is fast. There are very few clicks to get to full-text content. With LibXplore, we are unlikely to reconvert the Google enthusiasts.

3. Federated searching will present results to the end-user, sorted by relevancy, with duplicates records merged and independent of the sources from which they came.

   Reality: As illustrated in Figure 2, LibXplore presents users with a summary results page, with results in two columns (Found and Retrieved). The Found column gives the total results for each database, the Retrieved column gives a sub-set (up to 30) of records which will be merged into a single results list if the user proceeds to view the combined results. This subset of results is then sorted. The default sort for LibXplore is by "relevancy" but the concept of relevancy in a federated search environment bears little relationship to what the user expects. As mentioned above, since different databases provide different amounts of information in a typical record, the sorting algorithm ensures that records from one or two full text databases will always appear as the "most relevant" no matter what the topic, if these databases are included in the databases searched.

   Users can, if they wish, choose to re sort the results by date etc. and a library could choose to set date as the default sort order, in preference to relevancy. The Google (and Google Scholar) ranking algorithms work sufficiently well to ensure that users get good resources easily. They have little (nothing?) in common with the metasearching ranking algorithms.

If a user chooses to retrieve more records, the retrieved set is resorted and the same records will probably float to the top of the list. The concept of viewing some records and then having to look at the same records again mixed in with new records leaves much to be desired. Figures 5 and 6 illustrate that the same records appear at the top of the sorted list.

Figure 5: Results display, using relevancy ranking

Figure 6: Results display when additional records are retrieved

In the example illustrated in Figure 7, which uses the Biological Sciences quickset to search for articles on “stem cell research”, the results from Expanded Academic once again float to the top of the list, using the ranking algorithm, and articles from People Weekly and US Newswire appear at the top of the list.

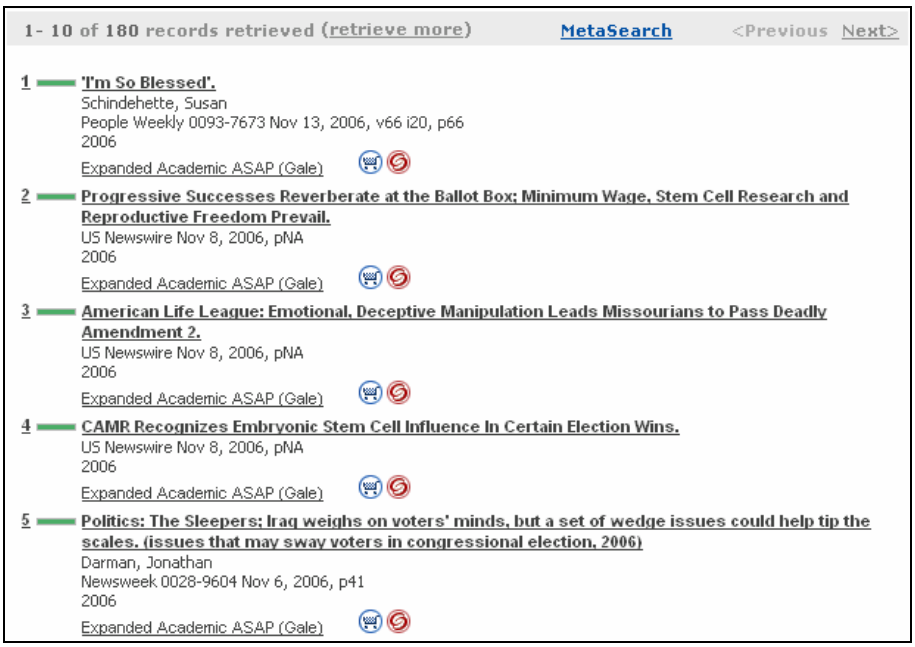

Figure 7: Results for the phrase “stem cell research” using the Biological Sciences quickset

The LibXplore results would be improved if Expanded Academic is excluded from the quickset list

for Biological Sciences but, at the time of writing, Expanded Academic is part of the quickset for Biological Sciences. Students (and their lecturers) are unlikely to value a system that produces articles from People Weekly at the top of the list.

For topics where there is an overwhelming amount of information, metasearching does not offer a useful strategy for locating quality information. Telling users that there are 5224 records on stem cell research and giving them the titles only of the first 150 (in screens of 10) doesn't really help, especially if predominantly non-scholarly materials appear as the most relevant.

One of the further claims of federated searching was that duplicates would be identified, and eliminated but, as Figure 8 illustrates, duplicate records are sometimes identified and sometimes slip through.


Table View Brief View Full View Jump to #: Sort by: Rank

1- 3 of first 3 records (combine more) <Previous Next>

No.	Rank	Author	Title	Year	Database	Action
1		Givoni, Sharon	Half a glass full of purple: [The Federal Court' decision in Cadbury Schweppes Pty Ltd v Darrell Lea Chocolate Shops Pty Lt	2006	APAFT (Informit)	
2		GIVONI Sharon	Half a glass full of purple.	2006	AGIS Plus Text (Informit) SNIPER (Informit)	
			Duplicate item - see # 2			
3		Givoni, Sharon	Half a glass full of purple	2006	SNIPER (Informit) AGIS Plus Text (Informit)	

Figure 8: Dealing with duplicates

To summarise, the metasearch promises remain, to a very great extent, unrealised. If we are attempting to win back the Google generation we still have a good way to go, and we may not be successful. If metasearching represents an attempt to develop a Google-like solution, the developers have, I think, misjudged the qualities that Google users value, or found them impossible to successfully replicate.

The enhancements offered by vendor platform to take account of feedback from users

At the same time as metasearching systems have been implemented, the database vendors have improved their database search interfaces beyond recognition. As Frost has argued,

"Database vendors have exerted considerable effort to improve their interfaces and search capabilities over the past several years. The costs for those improvements must be passed on to the customers, the libraries. Should libraries now avoid using the improvements for which they have paid?" (Frost 2004) p. 68

Users are typically offered the following options via the native interface for online databases"

1) A simple search option.

   This is usually the default option, but a library can often choose whether the simple or advanced search option is the default option. Database vendors will often allow parameters to be set locally, allowing a library to take account of the demands and degree of sophistication of their user population.

2) An advanced search option.

   The advanced search option allows the user to build complex searches – to search by author, for words in the document title, words in the publication title and to search for words anywhere in the full-text, when full text documents are included in the database. As Figures 9 and 10 indicate, the advanced search options are often numerous. Compare this with the Advanced search options in LibXplore (Figure 11).

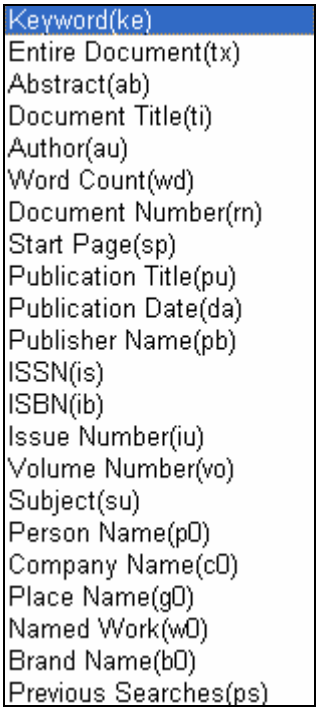

Figure 9: Field search options in Expanded Academic

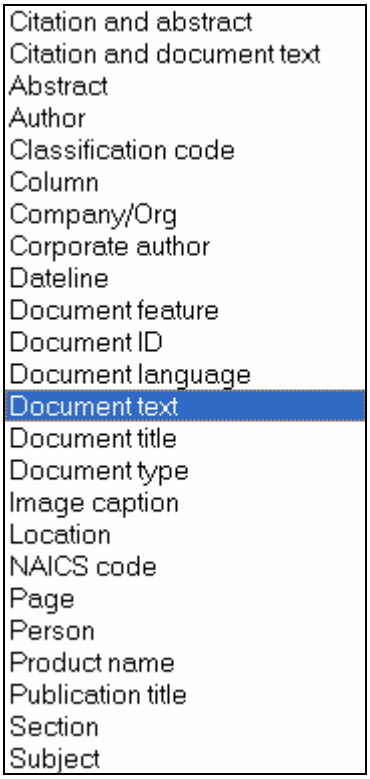

Figure 10: Field search options in ProQuest

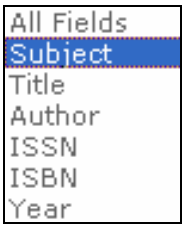

Figure 11: Advanced Search options in LibXplore

3) The ability to segment a database according to qualities of the documents contained therein. For example, ProQuest and Expanded Academic allow you restrict a search to documents with full text and/or peer-reviewed/scholarly publications.

   For databases with mixed content, this facility to exclude non-scholarly material is a key feature. As Gideon Haigh has noted (Haigh 2006), context and authority are concerns with documents sourced from the Internet and one of the shortcomings of the Internet is that this may not be obvious in the results of a search-engine generated search.

   The database vendors that include both popular and scholarly content have responded to concerns about authority and have developed ways to include/exclude popular content. In the metasearch systems that I've seen so far, this basic feature (like many other basic features) of the native search interface is lost in translation!

   Many other databases allow you to include or exclude particular types of documents. For example, Historical Abstracts (see Figure 12) includes four types of records (articles, books, collections and dissertations) so you can choose which collection(s) you search. And many databases allow you to include/exclude non-English language materials. The ability to make use of all of these limits is lost in the LibXplore environment.

Figure 12: Historical Abstracts

4) The ability to make use of subject headings in sophisticated ways is another enhancement offered in the native interface and lost in the LibXplore environment. In the ProQuest environment, once the user performs a search, you will see a list of suggested topics. These are alternate topics related to the search terms entered. Suggested Topics appear in order by relevance (best suggestions and matches first) and often contain pairs of index terms to help focus results. As Figure 13 illustrates, a search for the word “card fraud” gives a number of useful suggestions for other search terms.

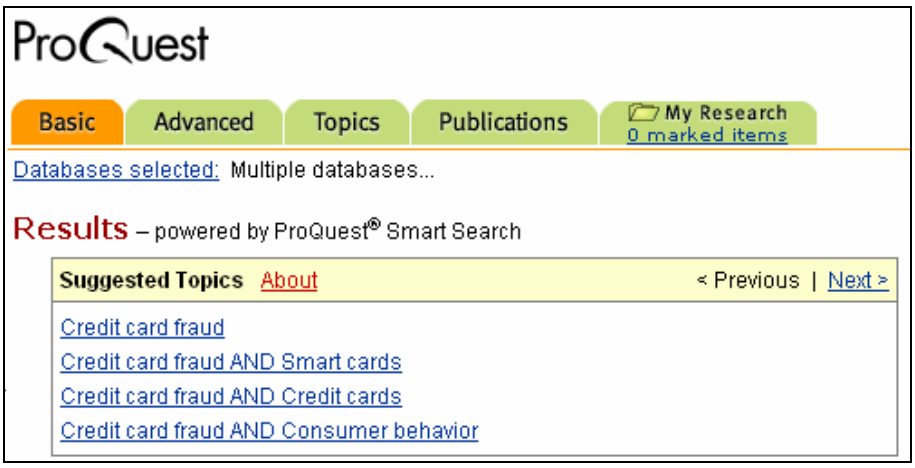

Figure 13: Suggested searches, based on the search terms entered

Many other database vendors offer similar features. LexisNexis has a “Suggest terms for my search” option. Based on the search terms entered, the user is offered a range of suggested terms (see Figure 14).

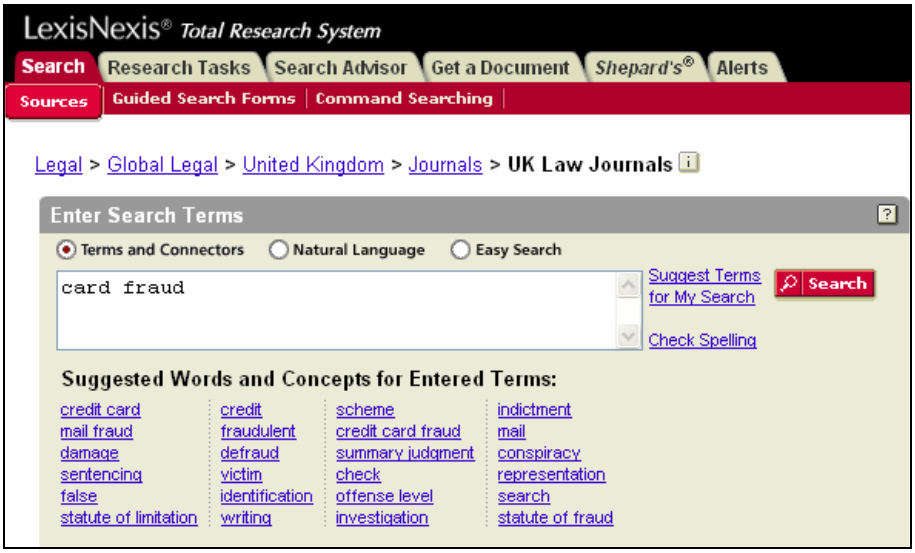

Figure 14: LexisNexis suggested words and concepts

In Medline and CINAHL, the facility to explode subject headings and select some or all subheadings

is a vital strategy in performing an effective search.

5) The provision of cross-searching via the vendor platform.

   Most vendors now enable cross database searching. For example, the Australian databases offered via Informit are cross-searchable and, via OVID, it is possible to cross-search AMED (Allied and Complementary Medicine), Biological Abstracts, CINAHL (Cumulative Index to Nursing & Allied Health Literature), Current Contents, EMBASE, Ovid MEDLINE and numerous other databases. The results are integrated into a single list, with a summary of results per database also available, and with the opportunity to remove duplicates. Figure 15 illustrates a search across four OVID databases, with results combined into a single list, and with the option of viewing records from specific databases.

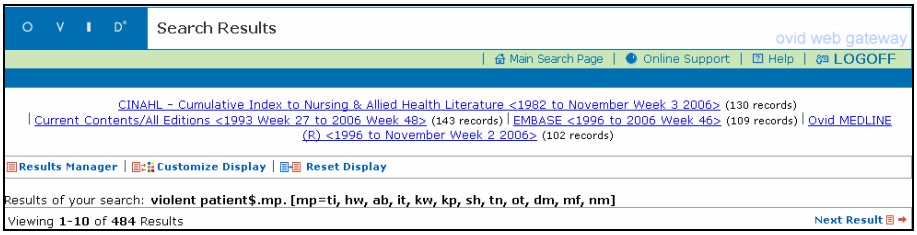

Figure 15: Ovid results display for a cross database search

6) Context sensitive help, and examples of how to construct search statements.

   It is common for the native interface to include context sensitive help and on-screen examples of relevant searches. The truncation character is usually visible in the native interface. Once again, this help information is lost in the LibXplore environment.

In summary then, each database interface takes account of the ways the people using that database will want to search AND the content of the database. Different scholarly disciplines have different requirements and the online vendors have done an excellent job of taking these requirements into account.

I would argue that a lot of learning has taken place and we are now at a stage when the search interfaces are simultaneously easy to use for the novice user and with a number of advanced options that relates to the relevant discipline(s).

At a time when universities are increasingly keen to prove that graduates have marketable skills (graduate attributes) and where many professions are moving to evidence based practice, it is vital that students graduate with high level research skills. They need to be effective searchers capable of devising sophisticated and comprehensive search strategies. They need to be able to use their discipline relevant database(s) in a competent manner, using the existing subject headings for that database. In a report on quality metrics, Chopra and Krowne have drawn attention to the need for searching to take account of disciplinary and sub-disciplinary vocabularies. They collected and analysed feedback from stakeholders in the scholarly community about the efficacy and value of key aspects of search technologies, including search interfaces, modalities, and results displays and they compared what users in the humanities expect compared with what users in the sciences expect. I would argue that the “one size fits all” approach taken in metasearching represents, in may ways, a step backwards rather than a step forward. To return to the plumbing analogy, the standard of plumbing in the metasearching environment is very basic.

The emergence of Google Scholar (and other ‘mega’ databases)

Cross-disciplinary databases also offer levels of sophistication as yet unmatched by metasearch systems. The free databases (Google Scholar and Windows Live Academic) and the multidisciplinary databases (Scopus, Web of Science) have emerged as strong competitors for metasearch systems and have quickly won end user loyalty. While it is true that these systems have (mostly) developed some time after metasearching was first proposed, they have quickly won ground. For those students simply wanting a

starting point, they would appear to be very good products. If near enough is good enough then why not make use of these systems for students pressed for time and wanting a few basic papers. Google and Google Scholar work fine in this situation and using Google Scholar avoids the considerable costs associated with building and maintaining local metasearch systems.

Google Scholar uses the number of times a paper has been cited as part of the algorithm to rank the results (see Figure 16). While this has some limitations, it is far preferable to the simple word counting algorithm used to determine relevancy in the metasearching environment. A search for “stem cell research”, conducted in the Biology, Life Sciences, and Environmental Science subject areas via Google Scholar produces more useful results than the same search via LibXplore, using the Biological Sciences quickset (see Figure 7).

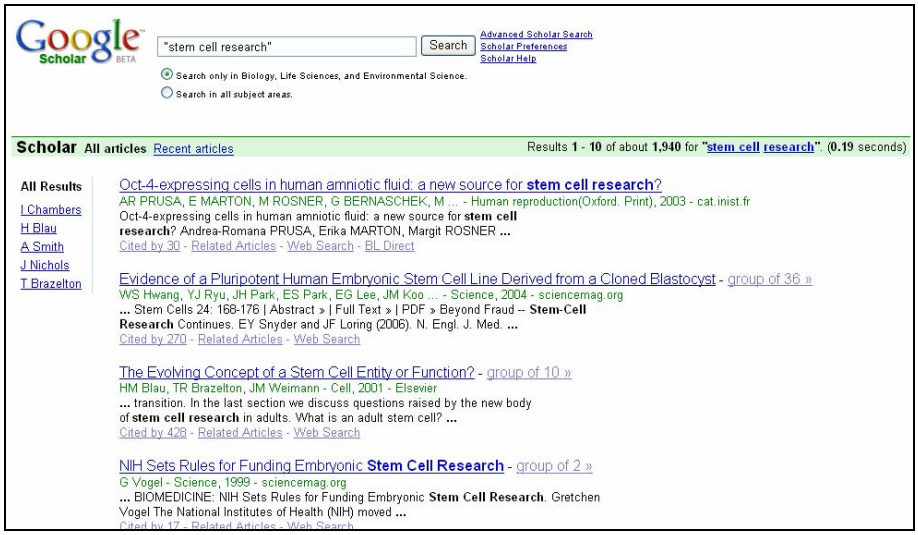

Figure 16: Results for the phrase “stem cell research” using Google Scholar

Other large databases can also work much more effectively as a starting point. For example, ProQuest, Expanded Academic and Scopus all function as excellent starting points for many research topics.

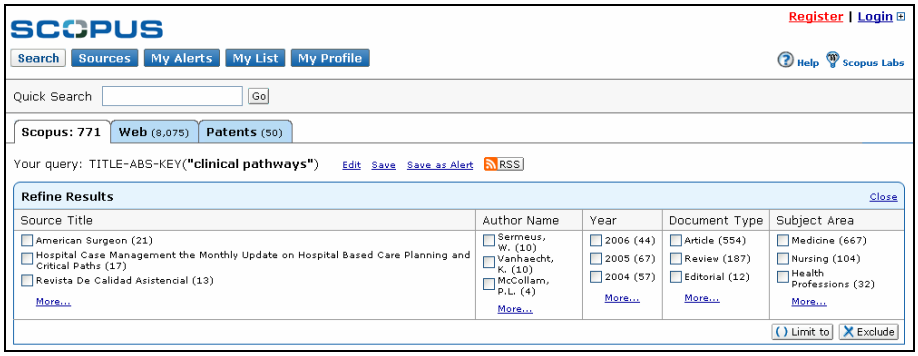

Figure 17: Clustering of results in Scopus

As Figure 17 illustrates, multidisciplinary databases can offer clustered results. A simple phrase search for the term “clinical pathways” returns clustered results – clustered by document type and subject area (Medicine 667 hits, Nursing 104 hits, Health professions 32 hits). Most search terms have multiple and discipline specific meanings – rather than aggregating results, the large aggregator databases can “disaggregate” results, in meaningful ways.

In an attempt to improve the display of results in the metasearching environment, one of the vendors, Serials Solutions, has recently announced that they will incorporate Vivisimo results clustering capabilities into its Central Search product. Clustering involves finding common terms and users can look at results in aggregated clusters. (Rogers 2006). Announcing this development, JR Jenkins, Serials Solutions’ group product manager has said:

”The beauty of clustering is that the general searcher runs very simple phrase queries, and clustering takes that very general term and exposes it to all the specifics attached to that term, so that users can come to the

search engine having to know a lot less than an advanced searcher and still benefit" (as quoted in Rogers 2006, p. 29)

Conclusion

Rather than being the promised step forward, metasearching as currently implemented may well be a step backwards. As Bawden and Vilar have observed after reviewing the literature on user expectations, "studies show that many, perhaps most users, find that traditional library systems, even in digital form such as OPAC, are disappointing, frustrating, illogical, counterintuitive, and intimidating" (Bawden and Vilar 2006) p. 349. While there is clearly a development phase with any new technology, there also comes a point when libraries need to take stock and review if the system is meeting the needs of the stakeholders?

Some years ago, the analogy of "drinking from a fire hose" was used widely in relation to the concept of information overload. I would argue that our users want filtered water – they are young and savvy and don't want too much information. They want the plumbing to work, but more importantly, they want quality at the end. For whatever reason, they have high expectations. Metasearching gives them quantity but the quality is lost. In the end, any gains are at a huge price and, I would argue, we have lost everything in translation.

Bibliography

Bawden, D. and P. Vilar (2006). "Digital libraries: to meet or manage user expectations." Aslib Proceedings: New Information Perspectives **58**(4): 346-354.

Burke, L., E. Fisch, et al. (2004). Implementing AARLIN at La Trobe University with a focus on end user reception. Breaking boundaries: integration and interoperability: 12th VALA biennial conference and exhibition. Melbourne. 2 - 5 February, 2004. .

Frost, W. J. (2004). "Do we want or need metasearching?" Library Journal **129**(6): 68.

Gow, E. and K. Roberts (2003). The AARLIN project: integrating the research information infrastructure- a case study. World Library and Information Congress, 69th IFLA General Conference and Council. Berlin.

Haigh, G. (2006). Information Idol; How Google is making us stupid. The Monthly. **9:** 25-33.

Highsmith, A. L. and B. C. Ponsford (2006). "Notes on Metalib implementation at Texas A&M University." Serials Review **32**(3): 190-194.

Marshall, P., S. Herman, et al. (2006). "In Search of More Meaningful Search." Serials Review **32**(3): 172-180.

Miller, T. (2005). "In defense of stupid users." Library Journal **130**(5): 58.

Oberhelman, D. D. (2006). "The time machine: Federated searching today and tomorrow." Reference Reviews **20**(3): 6-8.

Rogers, M. (2006). "Serials solutions debuts Vivisimo; federated searching gets supercharged with results clustering technology." Library Journal **131**(18): 27-29.

Tallent, E. (2004). "Metasearching in Boston College Libraries." New Library World **105**(1196/1197): 69-75.