LOST IN TRANSLATION: THE REALITY OF FEDERATED SEARCHING[1]

Dennis Warren

Federated search systems promised to solve the problems presented by unique vendor-supplied search interfaces, but since federated searching was first proposed the goalposts have moved and the environment has changed considerably. This article discusses some of the changes that have occurred, such as 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 deliver what was promised.
AARL December 2007 vol 38 no 4 pp258-269

Dennis Warren, Law Librarian, La Trobe University, Victoria 3086. E-mail: d.warren@latrobe.edu.au

The development of federated searching, or metasearching, systems has become one of the phenomenal growth areas in academic libraries in the last five years. But what commenced as a simple idea – allowing the end user to search across a range of commercial databases simultaneously – has, in many ways, proved difficult to implement. Although it was easy enough to produce a single search box, it has proved difficult to produce meaningful results at the end of the search 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. To date, however, it has not been clear that the results of federated searching implementations are meeting the needs of end users.

Those libraries that have been willing to report on their implementation of federated searching applications have described missed deadlines, soft launches and compromises made along the way. Much of the effort has gone into what one commentator has called the 'plumbing layer', where time is spent building connections to diverse resources.[2] 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 system. They have noted that many library staff had 'higher pre-implementation expectations than the software could support' and that staff 'object[ed] to the way in which federated search systems 'dumb down' native interfaces'.[3]

This article explores the pre-implementation promises made for federated searching, and compares them with current experiences for the end user, using the La Trobe University federated search implementation as the basis for the comparison. It also compares the sophistication of native interfaces (that is, the search interfaces supplied by database vendors) with the unsophisticated nature of searching in the metasearch environment. In the implementation process, many compromises have been made, and the question has to be asked: is the current implementation of metasearching close enough to the original idea to merit 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?

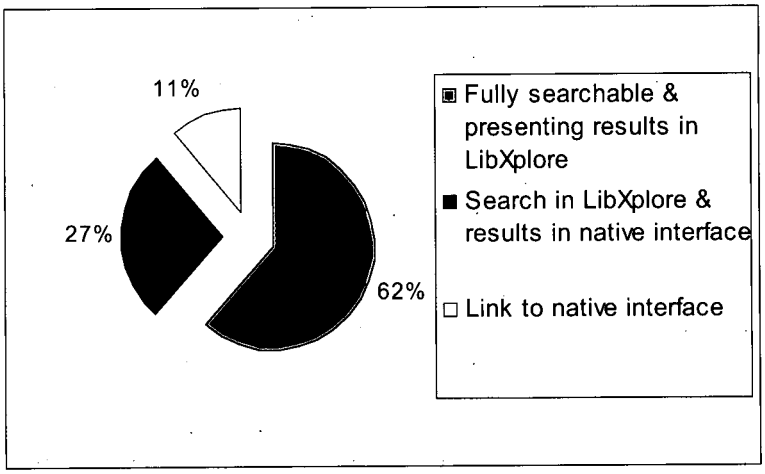
The examples used in this paper are drawn from La Trobe University's federated search system, known as LibXplore. La Trobe University and eleven other universities are part of the Australian Academic and Research Library Network (AARLIN) using MetaLib (version 3.13) from Ex Libris and the OpenURL link resolver SFX as their portal solution. The early phases of the AARLIN project have been documented elsewhere.[4] The combination of MetaLib and SFX has been widely adopted both in Australia and overseas as the preferred platform for the development of federated searching. A number of other Australian universities that are not part of the AARLIN consortium have independently used this combination of software to develop their own portal solutions.

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 generally fallen short of what was promised. It is possible to discern three broad promises or themes in the literature. These promises are outlined here and then compared with the extent to which they have been met via LibXplore.

The first main argument for federated searching is that it 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'.[5] However, federated searching via MetaLib allows the end user to search only a fixed number of databases simultaneously. As Figure 1 indicates, of the active resources in LibXplore, only 62% are fully searchable and capable of presenting search 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



Whilst users can build their own set of databases to search (up to 15), novice users are expected to make use of 'quicksets', or pre-built topical collections of databases. Figure 2 illustrates a search for the phrase 'credit card fraud', using the Law quickset, which is made up of five databases.

Figure 2: Search Using the Law + Legal Studies Quickset

QuickSearch Results

Search for "credit card fraud" in "Law + Legal Studies"

View retrieved Cancel

Database Name	Status	Found	Retrieved
CINCH (Informit)		106	30
Expanded Academic ASAP (Gale)		1678	30
AGIS Plus Text (Informit)		13	13
PsycINFO 1985- (Ovid)		3	3
APAFT (Informit)		9	9

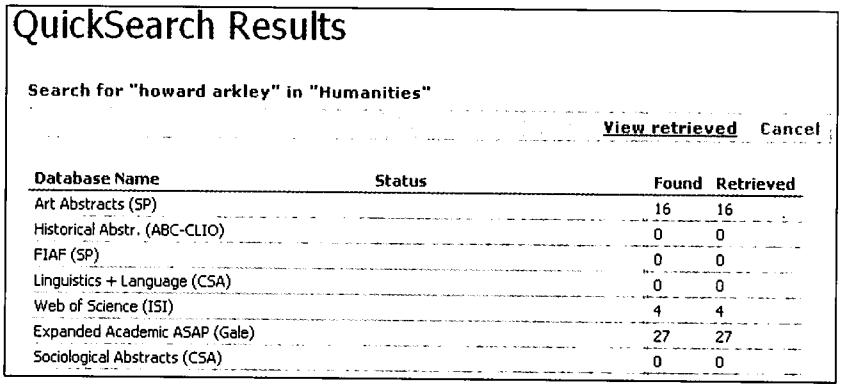
As the above search illustrates, the quicksets are often made up of both full-text databases (Expanded Academic ASAP) and indexes. When a full-text collection is included in a set of databases to be searched, the results from the indexes will be less numerous and the ranking algorithm will ensure that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

With the presentation of a pre-built subject quickset to users comes the risk that end users will not make use of the most appropriate databases for particular topics. 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, CRInfo (the Conflict Resolution Information Source), a free database that is tailored for the topic, would have to be included. For intellectual

property law, one of the best databases is Sniper, and for the credit card fraud example in Figure 2, ProQuest would have some value as it is certainly as relevant as Expanded Academic ASAP. 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 set limit of 15 databases, so metasearching does not really solve the problem of getting the end user to the most relevant databases. As a standalone resource discovery tool metasearching is no better than the long alphabetical lists arranged by topics that libraries offered previously. There are almost 300 databases configured for LibXplore. How can we be sure that users get the right ones?

To emphasise this point further, does this environment make it easier for the art historian or student of art history to find the art history databases? The answer is clearly no. If the end user knows that art history falls within the domain of the humanities (and there would be some art history students who do not see this connection) and selects the Humanities quickset, their topic will be 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? Would it not be more relevant to include the Bibliography of the History of Art in the quickset?

Figure 3: Search Using the Humanities Quickset


QuickSearch Results

Search for "howard arkley" in "Humanities"

View retrieved Cancel

Database Name	Status	Found	Retrieved
Art Abstracts (SP)		16	16
Historical Abstr. (ABC-CLIO)		0	0
FIAF (SP)		0	0
Linguistics + Language (CSA)		0	0
Web of Science (ISI)		4	4
Expanded Academic ASAP (Gale)		27	27
Sociological Abstracts (CSA)		0	0

Is the solution 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,

> students 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.[6]

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 if a user tries this approach he or she is then presented with a list of 16 'resources' and no obvious way to form a subset to search. 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. 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.

Figure 4: Art History 'Resources'

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	

As Frost has argued, improving the beginning researcher's selection of research tools is part of the educational process. Rather than attempting to make the underlying databases transparent, we should be encouraging selection of the most appropriate tool for the task. 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.[7]

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

The second main argument for federated searching is that patrons have stopped using resources because they are frustrated by the number of dissimilar search interfaces they must use to access database content.[8] This is an argument about redesigning the interface for access to databases which runs along the

following lines: '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'.[9] According to the proponents of this view, since federated searching eliminates dissimilar search interfaces, metasearching will entice a generation of Google fans back to quality library databases. Putting this line of thought succinctly, Marshall, Herman and Rajan have argued that 'federated search provides a way for librarians to reclaim the community that Google snapped up'.[10]

But do we have evidence that end users are 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[11] for a comprehensive review), there is also some anecdotal evidence that usage of databases in libraries was going up, not down, prior to the introduction of metasearching. And 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. The reality (explored below) is that the user interfaces have become much better. Indeed, by comparison with the federated search 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 subscription library databases.

To what extent does metasearching offer end users the Google experience? In many ways federated searching has more in common with traditional database searching than with Google. If two words are entered, 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. Using LibXplore, it is necessary to add an asterisk to activate truncation, but there is no onscreen help telling 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 differ. Google allows for spelling errors and suggests preferred spelling, through its 'Did you mean?' functionality. Google does not require the user to select resources to search. Google is fast and very few clicks are required to get to full-text content. With LibXplore, we are unlikely to reconvert Google enthusiasts.

The third argument for federated searching is that it will present results to the end user sorted by relevance, with duplicate records merged and independent of the sources they are drawn from.

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 subset of up to 30 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 relevance, but the concept of relevance in a federated search environment bears little relationship to what the user expects. Claims that results are being ranked are perhaps misleading; what is being ranked is the first results set returned by the search, not the entire search results. As noted 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 or other characteristics, and a library could choose to set date as the default sort order in preference to relevance. The Google (and Google Scholar) ranking algorithms work sufficiently well to ensure that users get good resources easily, on the first screen. But the same cannot be said of the metasearch ranking algorithms.

For topics where there is an overwhelmingly large amount of information, metasearching does not offer a useful strategy for locating quality information. Telling users that there are 5 224 records on stem cell research and giving them the titles of only the first 150 (ten per screen) doesn't really help, especially if predominantly non-scholarly materials appear as the most relevant. Some vendors have claimed that clustering results will solve this problem, but, once again, what is being clustered is simply the first results set returned by the search and not the entire search result.

A further claim of federated searching was that duplicates would be identified and eliminated but, as Hane suggests, true de-duplication is virtually impossible: 'Vendors that claim to do de-duping usually are just de-duping the first results set returned by the search'.[12]

To summarise, the promises of metasearch systems 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 either misjudged the qualities that Google users value or found them impossible to replicate successfully.

ENHANCEMENTS TO THE NATIVE SEARCH EXPERIENCE

At the same time as federated systems have been implemented, database vendors have improved their database search interfaces dramatically. 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?[13]

Users are typically offered the following options in 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 words anywhere in the full text, when full-text documents are included in the database. As Figure 5 indicates, the advanced search options are often numerous. Compare this with the advanced search options in LibXplore (Figure 6).

Figure 5: Field Search Options in Expanded Academic

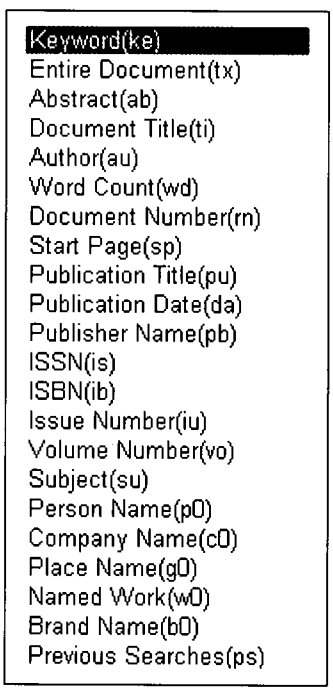
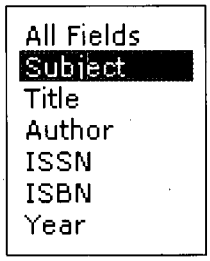


Figure 6: Advanced Search Options in LibXplore

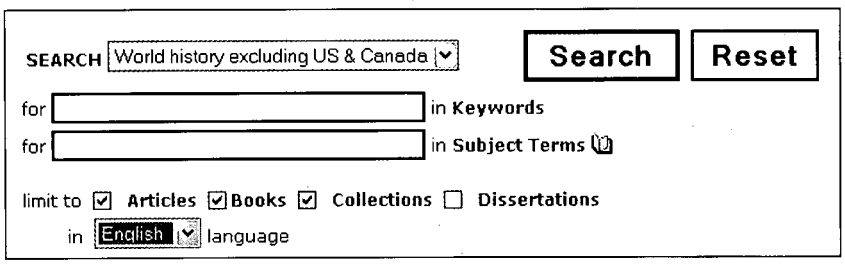


3 *The ability to segment a database according to qualities of the documents contained therein*. For example, ProQuest and Expanded Academic allow users to restrict a search to documents with full text and/or to peer-reviewed/scholarly publications.

For databases with mixed content, the facility to exclude non-scholarly material is a key feature. As Gideon Haigh has noted,[14] context and authority are concerns with documents sourced from the Internet and one of the shortcomings of the Internet is that context and authority may not be obvious in the results of a search-engine generated search. The vendors of databases that include both popular and scholarly content have responded to concerns about authority and have developed ways to include or exclude popular content. In the metasearch systems developed so far, this basic feature is one of many basic features of the native search interface that are completely lost in translation.

Many other databases allow the user to include or exclude particular types of documents. For example, Historical Abstracts (see Figure 7) includes four types of records (articles, books, collections and dissertations) so the user can select which collection is searched. Many databases allow the user to include or exclude non-English language materials. The ability to make use of any of these limits disappears in the federated searching environment.

Figure 7: Historical Abstracts



The ability to make use of subject headings in sophisticated ways is another enhancement offered in native interfaces that is not available

in the federated searching environment. In ProQuest when the user performs a search, a list of suggested topics is presented. These are alternative topics related to the search terms entered. Suggested topics appear in order of relevance (best suggestions and matches first) and often contain pairs of index terms to help focus results. Many other database vendors offer similar features. Based on the search terms entered, the user is offered a range of suggested terms. In Medline and CINAHL (Cumulative Index to Nursing & Allied Health Literature) the facility to explode subject headings and select some or all subheadings is a vital strategy in performing an effective search.

4 *The provision of cross-searching via the vendor platform.* Most vendors now enable cross-database searching. For example, the Australian databases offered through Informit are cross-searchable, and in OVID it is possible to cross-search AMED (Allied and Complementary Medicine), Biological Abstracts, CINAHL, 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.

5 *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 often lacking in the federated searching environment.

In summary then, each database interface takes account both 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 vendors of online databases have done an excellent job of taking these requirements into account. 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 offer a number of advanced options that relevant to different disciplines.

At a time when universities are increasingly keen to prove that graduates have marketable skills (graduate attributes) and 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 databases in a competent manner, using the existing subject headings for that database. In a report on quality metrics, Chopra and Krowne[15] have drawn attention to the need for searching to take account of disciplinary and subdisciplinary 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 with what users in the sciences expect.

The 'one size fits all' approach taken in the federated searching environment represents, in many ways, a step backwards rather than a step forward. Returning 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 federated search systems and have quickly won end-user loyalty. While it is true that these systems have mostly developed some time after federated searching 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 well 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. While this has some limitations, it is far preferable to the simple word counting algorithm used to determine relevance in the federated searching environment, especially given that the size of records varies widely from one database to another.

CONCLUSION

Rather than being the promised step forward, federated searching 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'.[16] Although there is much interest in a solution to this problem, it is not clear that we have found one yet. With any new technology there is a development phase, but there also comes a point when libraries need to take stock and review whether system developments are meeting the needs of the stakeholders. Federated searching is still a long way from delivering the hoped-for seamless cross-database access to the scholarly literature.

NOTES

1. An earlier version of this paper was presented at Information Online 2007 13th Exhibition and Conference Sydney 30 January-1 February 2007. The views expressed here are the author's personal views and not the views of La Trobe University Library.
2. Jenkins cited in M Rogers 'Serials Solutions Debuts Vivisimo: Federated Searching gets Supercharged with Results Clustering Technology' *Library Journal* 2006 vol 131 no 18 pp27-29.
3. A L Highsmith and B C Ponsford 'Notes on MetaLib Implementation at Texas A&M University' *Serials Review* 2006 vol 32 no 3 pp190-194.
4. E Gow and K Roberts 'The AARLIN Project: Integrating the Research Information Infrastructure – a Case Study' *World Library and Information Congress 69th IFLA General Conference and Council* Berlin 2003; L Burke et al '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.
5. D D Oberhelman 'The Time Machine: Federated Searching Today and Tomorrow' *Reference Reviews* 2006 vol 20 no 3 p6.
6. E Tallent 'Metasearching in Boston College Libraries' *New Library World* 2004 vol 105 no 1196/1197 p74.
7. W J Frost 'Do we Want or Need Metasearching?' *Library Journal* 2004 vol 129 no 6 p68.
8. P Marshall, S Herman and S Rajan 'In Search of More Meaningful Search' *Serials Review* 2006 vol 32 no 3 p173.
9. T Miller 'In Defense of Stupid Users' *Library Journal* 2005 vol 130 no 5 p58.
10. Marshall, Herman and Rajan op cit p180.
11. D Bawden and P Vilar 'Digital Libraries: to Meet or Manage User Expectations' *Aslib Proceedings: New Information Perspectives* 2006 vol 58 no 4 pp346-354.
12. P J Hane 'The Truth about Federated Searching' *Information Today* 2003 October p24.
13. Frost op cit p68.
14. G Haigh 'Information Idol: How Google is Making us Stupid' *The Monthly* 2006 no 9 pp25-33.
15. R Chopra and A Krowne 'Disciplining Search/Searching Disciplines: Perspectives from Academic Communities on Metasearch Quality Indicators' *First Monday* 2006 vol 11 no 8 at http://firstmonday.org/issues/issue11_8/chopra/index.html.
16. Bawden and Vilar op cit p349.