Version : Accepted Manuscript (January 2006)

Published in *Journal of Electronic Publishing,* vol. 9, no. 1, Winter 2006; Ann Arbor Michigan, Scholarly Publishing Office, University of Michigan, University Library

GOOGLE SCHOLAR: POTENTIALLY GOOD FOR USERS OF ACADEMIC INFORMATION

by Frederick J. Friend

Abstract
Use of the Google search engine is commonplace amongst all sectors of the academic community. The development of the specialist Google Scholar search service will benefit the academic community in bringing to their attention content more relevant to their needs. The vast number of Web sites containing potentially relevant information requires a search engine ranging over many millions of Web sites but with the ability to target very specific types of information. The Google Scholar service has the potential to grow if it develops close contacts with both providers and users of academic information. Use of Google Scholar will benefit the authors and managers of open access content, but there are opportunities for all types of academic content providers in the way Google Scholar is set up. Google Scholar will face competition and have to keep pace with user expectations and technological developments.

Google Scholar: the specialist shop in the Google shopping mall.

The extent to which the use of the Google[1] search engine has permeated academic culture is illustrated by a single statistic from a recent survey of academic authors: 72% of authors are using the Google search engine to search for scholarly articles[2]. We can easily believe that *students* would take to Google like ducks to water. It is easy to use, saves them time in writing essays, and links to a vast amount of information. That students would use Google fits with our prejudice about young people always wanting an easy solution without too much work, and using their familiarity with the new technology to achieve that result. Now it is out in the open that those of us in the academic world who have passed the student phase in our lives also use Google extensively! And we use Google for exactly the same reasons as students: ease of use, saving time, and access to a

[1] Information about Google is available on the Google international web-site http://www.google.com/intl/en/about.html

[2] Alma Swan and Sheridan Brown *Open access self-archiving: an author study* May 2005 http://www.keyperspectives.co.uk/OpenAccessArchive/2005_Open_Access_Report.pdf

wide range of resources. This revolution in the behaviour of people from every academic generation has crept up on us.

Valuable though the main Google search engine is, its limitations for anybody searching for academic content are readily observable. A search may show up many references of no academic value at all, not necessarily because the quality of the content to which a link is provided is poor but because the content is not of relevance to learning or research. Words used in searching often have different connotations, and a search may reveal content related to a different meaning of a word, although in practice such situations are rare. More common will be inadequate search results due to the failure of the search engine's inability to recognise the context of the words used for searching, the classic example of which is "lead", which is both a verb and a noun and has different connotations when talking about pencils and about chemistry. Also common are human errors in searching, particularly a failure to be precise in listing search terms. Google and other search engines are very sophisticated and the more we use their sophistication, the better the results we achieve. As an illustration of this point, a Google search under the words "open access" revealed 598,000,000 entries, nine of the first ten of which would have been useful to anybody investigating open access to research outputs, but beyond the first ten results the reader would be looking for the proverbial needle in a haystack. Using the "Advanced Search" facility on Google did reduce the size of the haystack in this search, providing many more than the nine immediate links in the general search, particularly through the use of the "all in title" facility and restricting the search to web-pages updated in the past six months.

An important limitation on use of a general search engine is the growth of content on the World Wide Web. A Google search at the time of writing this article did not reveal any recent estimate of the number of Web sites world-wide, but the estimate of the total number of domain names is over 45 million (http://www.zooknic.com/Domains/counts.html), and while some domains do not have Web sites, many others (such as those at universities) have numerous different sites under their domain. That creates real problems for users seeking the information most relevant to their needs. A reader with very little knowledge of a subject is easily led down paths into an information quagmire when so many Web sites are available. So the development of specialist search services, such as Google Scholar, is an inevitable and welcome development. The scope of Google Scholar is defined on the web-site[3] as: "Google Scholar enables you to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research. Use Google Scholar to find articles from a wide variety of academic publishers, professional societies, preprint repositories and universities, as well as scholarly articles available across the web." Where Google Scholar potentially has an edge over some other services is that it combines the specialist approach with its existing advantages. When we shop we like to combine our search for a specific product in a specialist shop with the advantage of a wide range of products in a huge supermarket or shopping mall. Google is the supermarket or shopping mall of information; Google Scholar is the specialist shop.

[3] http://scholar.google.com/scholar/about.html

The limitations of Google Scholar as it currently exists, however, can be illustrated by continuing the example of a search under the words "open access". A Google Scholar search did (at the time of writing) reduce the number of results from 598 million in Google to 1,250,000 in Google Scholar, and using the "Advanced Search" to limit the results to articles between 2000 and 2005 reduced the number to 28,000. However, the Google Scholar search would at this time have been less useful to the reader seeking an overview of the topic, as some general Web sites on open access were not revealed. On the other hand, more specialist journal articles were higher on the list, which would be useful to the reader who already has an understanding of the topic. This example may illustrate the fact that when we are not sure about what we want to buy, we do best to go to a shopping mall; when we know we want to buy a specific product we go directly to a specialist shop. The example may also illustrate the need for further development of Google Scholar.

Swimming in the sea of information

Users of academic resources come to a search service with a spectrum of knowledge about the topic on which they wish to gain more information and with a variety of techniques they have learned through previous experience of searching databases. They may, for example, know that a particular author has written upon the topic in which they are interested and will do no more than a simple author search. More commonly they will have in their minds a number of words which relate to the subject of their search, words which may or may not correlate to the words an expert in the subject or an information professional would use. In order to make a successful search out of this range of knowledge and awareness - from vague to precise, from novice to expert - users need key-word subject searching to the whole corpus of academic literature, both subscription-based and open access content. A great variety of search services already exist but most are limited in their scope or not user-friendly enough to be used without professional guidance. Some electronic reference services are available for specific enquiries. OCLC's "QuestionPoint"[4] for instance uses the collaborative knowledge of reference librarians. But for day-to-day information retrieval users need a service in which they can undertake their own searching of full-text databases and then select the text they recognise as being relevant to their needs.

Various portals, some providing links to resources on specific subjects, lead users to relevant content but the texts users are able to reach through the portal is often very limited. Even if relevant full-text is available, the number of clicks on the mouse that users have to make to get to the full-text tries the patience of the average user. For example, the British Academy Portal[5] has links to a very wide range of humanities content, a good resource. A student told by a lecturer that Philip Melanchthon's "History of the life and acts of Luther" was available through that Web site would have to make a minimum of five clicks on the mouse to reach the text, and more if the student guessed

[4] http://www.oclc.org/questionpoint/about/default.htm
[5] http://www.britac.ac.uk/portal/

incorrectly which resource to go to at one stage in the search. The challenge for information professionals and for service providers is to use the power of the Internet to enable the user to make a comprehensive search and reach the full-text of a small selection of relevant content within three or four clicks of the mouse? The technical means to provide a comprehensive searching and linking service for academic use are already in place, through the use of the DOI and the Open URL standard but no current service has yet been able to meet the easy and comprehensive approach offered by Google.

In times past students or young researchers reached the full-text of relevant journal articles by first consulting the holdings of their local library, following citations in journal articles they found locally, and then applying on inter-library loan for those articles. That method served students and researchers well for many generations but it is proving very inadequate as the holdings of academic libraries decline. Electronic document delivery is faster than paper inter-library loan but still too cumbersome and expensive to be used as a large-scale substitute for local holdings. The contrast between the limited information available in their local library and the *unlimited* information available on the Web is not lost on a generation of students that has already been brought up in an Internet world. Information users further along the career road may be more familiar with sources in their field, but the statistic quoted above[6] shows that Google is valuable even to those who have become authors. The academic world does not consist of a series of isolated subject communities, and as soon as authors begin to look for relationships between their work and the work of other authors in different disciplines, the need for search engines covering content across a wide range of disciplines becomes apparent. Although the feeling that each academic discipline is self-contained still exists in the minds of some faculty members, increasingly the links between very different disciplines – such as the link between medicine and ethics or sociology – are better recognized and bridged with search engines than with stand-alone journals. Some specialist databases may prove more useful than Google Scholar today, but this situation may change over time. Elsevier offer users an opportunity to compare search results from their Scirus database with results from Google (not Google Scholar) on scientific content[7]. Information searchers should accept the invitation to "do the Scirus-Google test", because it encourages competition between providers of information, which should lead to more cost-effective services.

The need to provide an effective link between user and content is to be found in the size and complexity of the sea of academic information. It is customary to speak of "a *pool* of information" and the image those words conjure up is of a small pool of information that is static, shallow, well-lit and confined. Rather what faces the user of electronic resources is a huge *sea* of information that is ever-moving, deep, dark and boundless. Into this sea the user has to plunge in order to find a quantity of information that is minute by comparison with the whole but of great importance to the user's need. In taking the plunge all that the user carries is a small bag of existing knowledge, perhaps a few words which describe the topic or an awareness of relationships, so that if the user comes close to the topic, its outline will be recognised by the shape of the information around it.

[6] See reference 2 above.

[7] http://www.extranet.elsevier.com/listman/scirus/nov04/google%20online.htm

Imagine yourself swimming in a sea at night and coming close to a shore-line whose outline you recognise as being close to where you wish to be. The opportunities for the intermediaries in the world of information to meet user needs lie in guiding the user to that small amount of content they will find valuable. We all need our personal lighthouse to guide us in the sea of information. Our preferences amongst databases often reflect our personal experiences in searching for information, how we search and what we find useful. Google is already well-established as a beacon guiding many individuals to the information they seek. It has a prominent position in the information landscape.

Future development of Google Scholar?

Given that Google is already used heavily by many members of the academic community, and given that it is developing services aimed specifically at that community, it seems likely that Google's influence upon the world of scholarly information will grow. How will growth in the use of Google services interact with growth in another part of the information world, namely open access to academic journal content? The development of both open access and a search service like Google Scholar have the potential to shake the foundations of the information world. The results from the general Google search engine already include open access content but the open access links are not clear to the reader and usually buried in a mountain of other links. Clearer identification of the status of items in lists of search results (for example identifying free content or content for which payment is required) would assist readers, and a more specialised service such as Google Scholar should be able to provide such version identification. It may be that the long-term commercial success of Google Scholar will depend on such value-added services.

It is in Google's interests to develop Google Scholar in ways which meet the needs of as many members of the academic community as possible. A universal approach has to remain Google's strength, and in order to achieve its goals Google will need the cooperation of the providers as well as users of academic information. Google is a commercial company with a high profile and its commercial success does not depend on the academic sector. However, a specialised service such as Google Scholar will survive or die according to its market success; the academic community may welcome its availability for a time but if it ceases to fulfil a valuable function users will shed no more tears at its passing than at the loss of any other information service that loses contact with its user community. Users of academic information services are not interested in Google Scholar as a business, but in the services it provides.

The development of open access repositories and journals is beginning to transform scholarly communication, and there could be pointers to success for Google Scholar in the way in which that transformation has happened. The strength of the open access movement is that it has arisen from and developed within the academic community it exists to serve. Open access services have advantages that commercial services do not. Because universities have committed time and money to establishing a repository, they are more committed to their success, and their repositories tend to have a greater rapport with the university communities than commercial services, which have a different part in the open access revolution as service providers to the academic community. Google is a commercial company, and if the services it provides to the academic community are to survive, the design of those services needs to be driven by the requirements of the academic community.

The challenge for Google will be to develop a rapport with the academic community, so that its services become as deeply embedded in the work of that community as open access repositories and journals are becoming. There is a commercial example for them to follow in the success of ISI[8], whose citation products are used in research assessment in many countries. ISI's dominance of this market has led to some criticism of its role, but as an example of a commercial company becoming a key player in academic life it is unrivalled. ISI has reached that key position by understanding the importance of quality issues to both academic and publishing communities, and using the contacts its staff have had with those communities to design services which benefit both. Google has a similar opportunity, through its understanding of the importance of searching Internet resources, but its staff need to develop close links with academic users if the opportunity is to be realised.

Building a better Google Scholar

Google Scholar, like other information providers on the Internet, needs to find solutions to a common problem in online searching, the provision of relevant, high-quality results. Today Google Scholar provides its own version of citation information, with a "ranking" technology that reports how often the item has been cited in other scholarly literature. This certainly tells the reader that authors have found this item valuable enough to cite, and tells authors how valuable their work has been to other authors.

All players - open-access and subscription publishers, managers of open-access repositories, and librarians - should collaborate to find a way to identify high academic quality in a search-engine world. Those measures may not be as dependent on the impact factors as they are in the print world; they may be new measurements that we need to identify. If Google Scholar is to better meet the needs of the academic community, it has to reveal to users the most relevant high-quality content from a vast potential source. Two apparently contradictory factors are important: having as many potential sources of information as possible but also having the ability to choose the best. Any restriction upon the size of the source of information will be bad for Google. Differentiation on the basis of quality of content needs to come through Google's search software, not through the size of the sea of information that forms its source.

Google Scholar needs to be able to search across as wide a range of high-quality academic content as is available on the Web, and authors and managers of open access content need their content to be found and used. These needs complement one another

[8] ISI's products and services are described at www.isinet.com

and are not in competition. One way to improve the provision of high-quality content is for Google scholar to include in its database all open access journals and all open access repositories at education and research institutions. The availability of open access content in repositories and open access journals adds considerably to the size of Google's source. While some open access content is already available through Google Scholar, the criteria for inclusion are not clear, and more types of open access content – in addition to journal articles – could make a search more valuable for a user. Content from university repositories and peer-reviewed open access journals is high-quality content, and it presents no barriers either in linking restrictions or in access by users. Open access content presents a search engine with low-hanging fruit for easy picking.

It is similarly in the interests of open access authors and the managers of open access content to encourage any means of finding the content. While general search engines already provide an excellent route for users to open access content, a specialised service such as Google Scholar has the potential to make life even easier for the user. The easier and more accurate a search service is for the user, the greater the use of open access content likely to result from a search. Open access also provides value to authors, because the lack of technical or payment barrier means greater usage and therefore higher citation levels.[9]

Much subscription content is also high quality, and may also present a search engine with low-hanging fruit in respect of ease of linking. Google Scholar now shows subscription content in its search results, and takes the user to the content only if the user's library has a subscription to the content. This arrangement will benefit subscription publishers when users of Google Scholar have a subscription to the content indexed. Interestingly it will also clearly benefit open access publishers. Users who wish to use a version of a work held in a repository will not lose out under this arrangement because a very helpful Google Scholar policy is also to group multiple versions of a work. Google Scholar's web-page on "Support for Publishers" describes how "Publisher's full-text, if indexed, is the primary version"[10] even as Google Scholar provides access to other versions. If several versions of a work appear on a search results screen and each version is clearly identified, users can choose the version that best suits their situation. This already happens with search engines when we choose not to select the first item on the results screen but go instead to the second or third item if the description of that item is closest to our requirements both in content and in format. Through such arrangements users of Google Scholar will have an advantage over users of search services that only search particular types of content (for example journal articles) or content from particular types of providers (for example commercial publishers). What this means for the reader or publishers is yet to be discovered.

Libraries and Google Scholar

[9] Research is still under way into the use and citation-levels of open access content, but a good description of the "effect of open access upon citation impact" is contained in the slide at http://www.ecs.soton.ac.uk/~harnad/Temp/OATAnew.pdf#search='Open%20access%20citation.

[10] http://scholar.google.com/scholar/publishers.html

The important role librarians play as intermediaries in access to academic information is recognised through Google Scholar's special services, which are explained on the Web site under "Support for Libraries"[11]. Two practical services are offered: a "link resolver" to library holdings and a link to OCLC's Open World database[12], both services intended to make it easier for users to locate in a library content identified in a Google Scholar search.

This is the kind of service Google Scholar needs to offer if it is to be a commercial success, embedded into information services in a similar way to ISI's citation services. The test of success for Google Scholar in winning the hearts and minds of the library community will lie in the way in which libraries incorporate Google Scholar into their own Web sites. For instance, the University of Texas Library has already added a special link to the Google web-site which enables users to trace content from a Google search in the University's own Web pages.[13] On the other hand some librarians may feel that a page with the text "Welcome to Google's university search of University of Texas Austin" gives readers an impression that one commercial supplier is favoured above others. The academic community may be wary of too close a relationship between publicly-funded and commercial services. Extra concern arises when the commercial supplier is in a monopoly position, or effectively dominates the market even if it is technically not a monopoly. It may be good for the academic community if Google Scholar has to face stronger competition than it appears to face currently.

The future?

Google Scholar has already raised much interest amongst the information community. The "chatter" on several e-mail lists is largely positive, particularly among open access supporters. Concerns and uncertainty remain, however, about Google's definition of "scholarly" in determining inclusion or exclusion, and also about the currency of the content. The Google Scholar database is not restricted to peer-reviewed content (a wise decision) but only time will tell whether either too much or too little useful content appears in Google Scholar search results. Google Scholar managers will have to fix their patchy currency because most users will require up-to-date information. These doubts about Google Scholar's value are capable of resolution, given close co-operation between Google Scholar managers and both users and providers of information.

One opportunity open to Google Scholar is to offer the academic community searches that recognise the context of the words used in searching. The lack of context-related searching forms the most significant weakness in the use of general search engines for academic purposes. If Google Scholar is to provide an effective context-related search service its designers have to be inside the minds of students and academic staff, thinking about words in the way they think, understanding relationships between words in the way

[11] http://scholar.google.com/scholar/libraries.html
[12] http://www.oclc.org/worldcat/about/default.htm
[13] http://www.google.com/univ/utexas

that fits with learning and research, knowing the context within which particular words are likely to be used. This is a considerable challenge, particularly for a world-wide service. Although the *academic* context of words crosses international borders, there will be differences in *cultural* context which will influence the information needs of students and academic staff. If a UK student reads a book or journal article written by a US author, the student will (assuming that the book or article is well-written) understand the academic context of the words used and will probably use the same words in a very similar local context, but behind some of the words there will be a US cultural context of which a UK reader may not be aware. Even more marked will be the cultural differences for users of Google Scholar coming from a non-Anglo-Saxon background. Such differences will affect the satisfaction of users of Google Scholar with the search results the service provides.

A further factor affecting the future for Google Scholar is the extent to which it may face competition. Google Scholar's potential competitors may be either commercial or public-sector services. Any such service requires a substantial financial commitment, which can often only be met on an on-going basis (by contrast with set-up funding) through commercial involvement. On the other hand public-sector research may already be providing the technical basis for an effective academic search service. Semantic Grid technologies[14] - "Grid" developments in academic computing combined with semantic web technologies - have the potential to transform any search for information. If Google is able to take such research developments and apply them before its potential competitors, it will be in a strong position, but equally the academic community itself may develop service applications based upon Semantic Grid technologies. What is certain is that the way in which information is sought and secured will within a few years be unrecognisable to users of today's systems.

Frederick J Friend
JISC Consultant
OSI Open Access Advocate
Honorary Director Scholarly Communication UCL
f.friend@ucl.ac.uk

[14] The semantic grid "vision", combining the semantic web with the Grid, is described at http://www.semanticgrid.org/vision.html