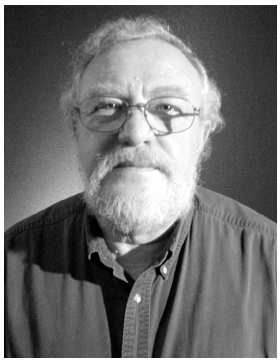
Key issue

Google Scholar



CHUCK HAMAKER ***BRAD SPRY***
Atkins Library
University of North Carolina Charlotte

First, remember anything we write about Google Scholar will likely be obsolete before it is published. Authors who complained about lack of date range indexing (Marshall Breeding's January 2005 Article in *Smart Libraries* 25:1, for example) were overtaken before their words saw the light of day. So all we can report in this brief review is what we can tell from using the service as of the end of December 2004.

Google Scholar apparently made a decision to index fairly completely all scholarly, known or cooperating publisher-based sites, but to only partially index university web sites based on file format identifiers, i.e. PDF or PS files. A secondary filter of some type limits most of those PDF and PS files to what seem to be journal articles. Standard bibliographic formats (citations) are probably recognized and their inclusion signals a scholarly article. Administrative notes, for example, posted as PDF files are generally excluded, perhaps because they often lack formal authors or footnotes. Of the 24,000 items at The University of North Carolina (UNC) Charlotte that a Google site search identifies when 'pdf' is used as a search term, fewer than 500 are identified in Google Scholar. This does not bode well for inclusion of special collections and other content being created by libraries specifically for the web. We could find nothing with the search 'Documenting the American South', for example. Another limitation is size of file, discussed by Gary Price at: http://blog.searchenginewatch.com/blog/041201-105511 (visited 3 January 2005).

Google Scholar is scholarly to a point, but what it does not include is important. If we understand correctly what it does index, it is time to get on with the much larger job of identifying more trusted scholarly sources. It has done a great job with the basic stuff, indexing 25% to 50% or more from many participating sites and obvious locations (such as arXiv) and identifying scholarly content through secondary means, i.e. citations and abstracting sources like PubMed and ACM. Between the first and third weeks of December, coverage tripled for many standard publishers. But inclusion based on a fairly limited primary source list and document format (or bibliographic citations) are just a beginning. Can it go beyond this to the rest of the scholarly resources on the web? Can it distinguish between a footnoted and non-footnoted scholarly item at a trusted source that is not a publisher?

They might consider it proprietary, but we really need to know – what is the list of completely indexed trusted sources? Can academia and publishers and librarians help expand the list of trusted sites or subsections of sites (as targets for complete indexing)? How? A way forward might be making public a list of trusted sources first (probably the CrossRef participating publishers is a core) and inviting some library-related groups to contribute additional sources. Whether Google Scholar develops into a reliable tool or just a curiosity may well depend on the answers to such questions.

Some questions answered about Google Scholar

Can I actually get these articles from a wide variety of sources?

Yes. The spidering Google normally uses has been turned into some fairly sophisticated indexing and

parameter filtering so what you find is references primarily to published articles, often with citation links and with links to variant versions of the same article. Based on checking Google Scholar entries for traditional articles with '.edu' extensions on other 'versions' and other signals of availability, in some subject fields we believe as much as 20% of articles published in the last few years may have perfectly legal copies available to readers outside the journals themselves. But whether or not articles are actually available, they are widely identifiable. In identifying articles, at least in the fields we checked, Google Scholar did an excellent job of finding recent publications.

Hello. My name is Googlebot. *May I come in?*

Some publishers and institutions may be surprised to find their content has been indexed by Google. To create its index, Google commands an army of software-based robots to crawl as deep as access controls allow.

Googlebot, the official name of Google's robot, also has sophisticated document archival abilities. If left unchecked, Googlebot will automatically archive copies of institutional and publisher content. Googlebot's abilities include transformation of PDF and postscript files into HTML, and creating cached copies of full-text HTML pages.

Luckily for content providers, Googlebot is obedient; you just need to know the commands:

Googlebot (Google's Web Crawler):
http://www.google.com/bot.html
The Web Robots Pages:
http://www.robotstxt.org

Can I get access to articles in journals I subscribe to?

Generally the answer is yes if you are logged in to your authentication system. However, there seem to be exceptions depending on what site or server entry point the publisher has permitted Google to spider. We suspect this is due to the direction into the publisher's databases, or the root URL used for the spider. Some libraries in the UK have noted problems with Athens login authentication at some sites.

What else can I really get?

It varies enormously but there are some clues in the record that will often let you know whether you can get a copy of the article or not. First, if there is a hyperlink that says 'view as html' or 'cached' then you can probably retrieve the article. Second, look for the additional locations to see if there is an .edu or international university site. If there is, the article is likely to be available at that site. Economics literature is widely available from preprint and postprint sources, physics literature is often on arXiv and mathematics articles are widely available. Note the format of the article though, as many of the engineering and maths articles, for example, are in postscript. But some postscript format articles have been converted to HTML views. During the first week of December, 47 of the first 100 (of 53,400) articles at the site: Blackwell-synergy.com had links to additional .edu locations. Blackwell's also makes a significant number of articles available free at its site, most on delayed access. These articles are not marked in anyway to be recognizable to users but open when the PDF is clicked.

How current is the indexing?

To answer this question we looked at the most recent issues of nursing journals at the Blackwell's site. On average, Blackwell nursing journals were indexed on Google Scholar over two-and-a-half months before they show up in CINAHL. To address this gap created by manual indexing lag, EbscoHost has introduced a Pre-CINAHL, similar to pre-PubMed content many are familiar with.

How comprehensive is the indexing?

Google Scholar cautions that site indexing does not work well. But 'exact phrase' searching does indicate somewhat the range of available content. We used 'exact phrase' searching for sites in the following table.

Many articles from these publishers are represented in Google Scholar through secondary rather than primary entries, so even though these might seem like low numbers, the representation is much higher.

Site	Entries 1 December 2004	Entries 20 December 2004	Approximate coverage (estimate)	Notes
AIP.org	97,600	159,000		Authorized user had difficulty retrieving articles
Blackwell-synergy.com	53,400	208,000	50%	Authorized users OK Many articles open access but not marked as such
Elsevier.com	11,800	54,000		cached or HTML versions – open access or server error
Extenza-eps.com		27,200	100%	All articles indexed and linked. Authorized users OK
Ingenta.com	128,000	343,000	20%	Authorized users OK
Sciencedirect.com	NONE	NONE		Secondary indexing only from ancillary sources
Taylorandfrancis.metapress.com	39,100	72,800	20%	Authorized users OK
Wiley.com	70,600	224,000	33%	Authorized users OK

What about content from publishers who are not co-operating with Google Scholar?

Some publishers have not participated, so their content is only represented by inclusion in secondary targets, i.e. citations in indexed articles or stand-alone indexing sources like PubMed. Many journal articles can be represented solely by citations from other sources. In short, even if a site does not participate, many of its articles can be identified through the index.

Chuck Hamaker
Associate University Librarian Collections and Technical Services, Atkins Library
University of North Carolina Charlotte
Charlotte, NC 28223, USA
Tel: 704 687-2825
E-mail: cahamake@email.uncc.edu

Brad Spry
Library Webmaster, Atkins Library
University of North Carolina Charlotte