Start your engines

Google has launched another challenge to commercial search services — this time aimed at scientists. But is the new engine running as smoothly as its fans hope? **Jim Giles** investigates.

As an undergraduate in India in the mid-1980s, Anurag Acharya had to write letters to scientists when he could not find the papers he wanted. It is a memory that makes the softly spoken computer engineer laugh. Now working at Google, Acharya is creating a search tool that aims to be the first choice for everyone from Indian students to Iranian professors. "I want to make it the one place to go to for scholarly information across all languages and disciplines," he says. And that ambition, he freely admits, is "simple to state, but not to achieve".

For a member of the public seeking a one-off scholarly article, Google Scholar is ideal. It is free to access, and as easy to use as the main Google search engine (see 'Inside information', opposite). But for academics with access to dedicated library resources, why make the switch? Most scientists rely on tried and trusted favourites, including subject-specific databases such as the US National Institutes of Health's PubMed or the NASA Astrophysics Data System, to find papers.

Since its launch last November, Acharya's Scholar engine has delighted and infuriated in equal measure. One librarian has even begun a blog following the search engine's progress. Although there are no detailed studies, many librarians report that faculty members and students are beginning to use the search engine; some suspect that Scholar will replace more established, and more costly, search tools. Figures from academic publishers also suggest that use of Scholar is growing rapidly: it already directs more online traffic to *Nature* websites than any other multidisciplinary science search engine.

Thomas Mrsic-Flogel, a neuroscientist at the Max Plank Institute of Neurobiology in Martinsried, Germany, and a regular PubMed user, has started to use Scholar. He says he finds the engine useful when he is not quite sure what he is searching for. Search results include citation links to other articles, so he follows the links until he finds something interesting — a function that PubMed, which does not track citations, cannot provide. "I follow the citation trail and get to papers I hadn't expected," says Mrsic-Flogel. "I have found papers that way that I wouldn't have found otherwise."

This citation tracking puts Scholar in direct competition with the fee-based search engines marketed by traditional science publishers. Until Elsevier launched its search engine, Scopus, in 2004, Thomson Scientific's Web of Science had a monopoly on citation tracking. Citation counts allow researchers, institutes and journals to follow the impact of individual articles through time, leading to metrics, such as journal impact factors, that are the bane and blessing of many academic careers.

But unlike Scopus and Web of Science, Scholar does more than just search the peer-reviewed literature. Some users like the fact that Scholar searches lots of non-traditional sources, including preprint archives, conference proceedings and institutional repositories, often locating free versions of articles on author websites. This 'grey literature' is growing in importance but remains poorly defined. It is widely assumed that Google considers a source scholarly if it is cited by another scholarly resource — but as online publishing evolves, so may this definition. Advocates of greater access to the scientific literature hope that Scholar will encourage more researchers to deposit their articles in free online repositories.

But how well does Scholar actually work? Librarians who have run systematic

C. DARKIN
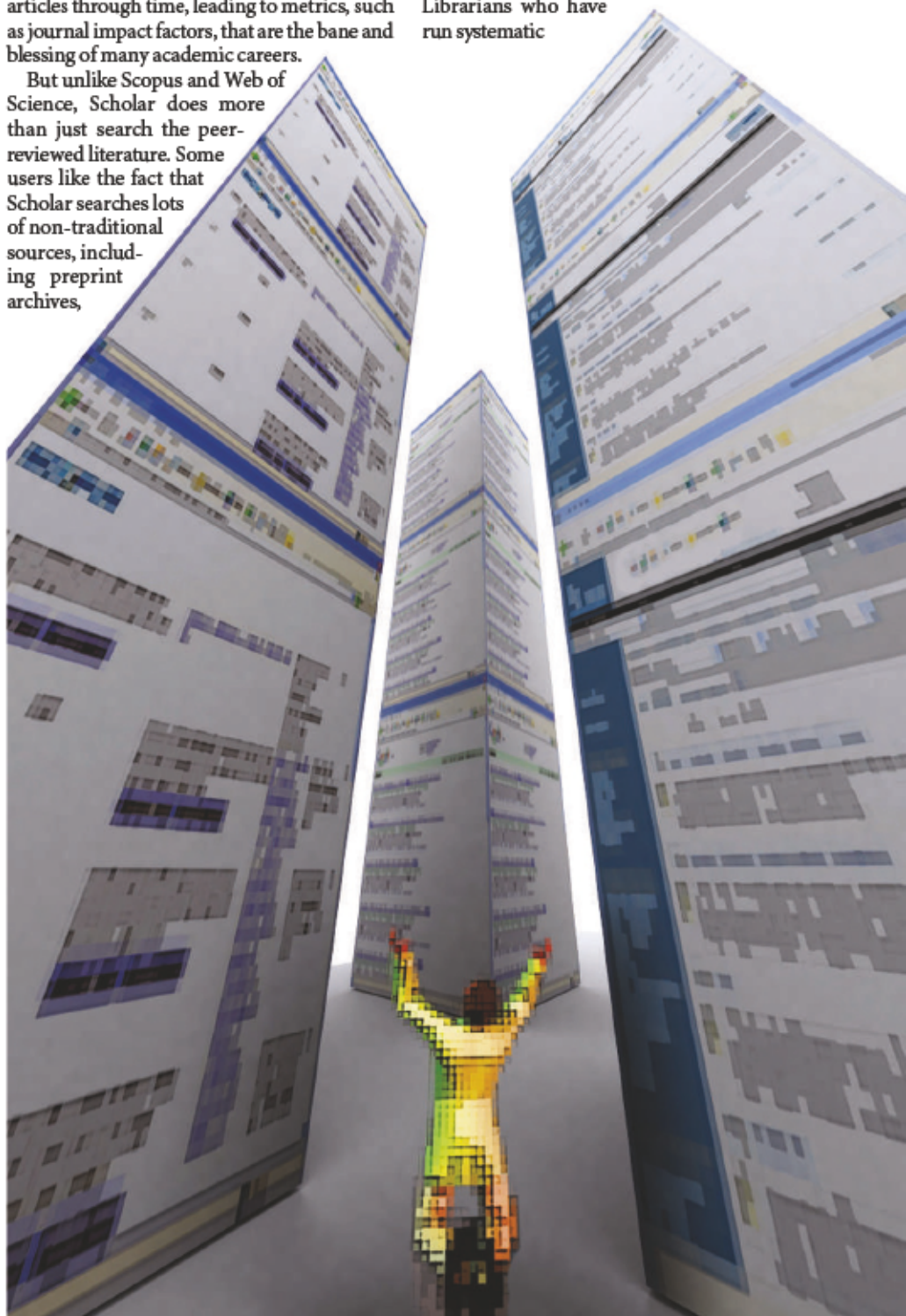
Inside information

Science search engines are fine for literature searches, but scientists inevitably need much broader information from the web. Searching using the main Google engine may take some coaxing, but a few tricks can help you to find the most relevant information faster, and to get a variety of views on a topic.

Google has advanced search options that will help you narrow your search, using more precise terms, or broaden it, using synonyms. Here we list some less well-known tips, using the drug Tamiflu as an example.

Site: Websites are often difficult to find your way around, so rather than wasting time endlessly clicking, just type 'site:' into your query followed by the website name. Searches can also be restricted to a domain name. For example, 'site:gov' will limit a search to US government sites, and 'site:nih.gov' to the National Institutes of Health. A search for Tamiflu at the World Health Organization, 'Tamiflu site:who.int', returns about 100 hits. A broader search, such as 'tamiflu site:edu', brings back more than 40,000 hits from US universities.

Filetype: A useful way to refine searches is to search for particular document types using the 'filetype:' query. A search for 'Tamiflu filetype:ppt' will return only PowerPoint presentations, which are usually conference talks. 'Filetype:doc' will often return project proposals or government texts, 'filetype:pdf' is more likely to return scientific information.

Define: This simple query will provide a definition of the words you enter after it, gathered from various online sources. The query 'define:Tamiflu' takes you to definitions in Wikipedia in several languages for example.

Quotation marks Ultimately, the web is about people, and if you are looking for contacts, or possible collaborators, there are some ways to Google scientists. The query, '"avian influenza" "workshop participants"', will bring back a few hundred hits, often with contact details for world experts among the top results. Variations of this will do the same in any scientific field.
Declan Butler

searches across several engines, say that Scholar performs well. A study published this year, which looked at more than 100 papers, concludes that Scholar finds similar numbers of citations to its commercial rivals[1]. Yet such results need to be interpreted cautiously, say information scientists. Critics point out that the study did not examine the list of citations to see whether they contained duplicated or erroneous entries.

A closer look at Scholar search results suggests that duplication may well be occurring. One of Scholar's harshest critics, Péter Jascó, an information scientist at the University of Hawaii in Honolulu, has taken the engine on numerous test drives. He has documented the results in unflattering terms on a website run by Thomson Scientific. In one extreme case, Jascó found that the first 100 results from a search for documents on 'computers' and 'intractability' returned 92 slightly different citations of a book entitled *Computers and Intractability* and only 8 other unique results.

Cite unseen

The source of this problem is the way in which Google adds records to its scholarly index. At Web of Science and Scopus, staff scan in the abstracts and references from print journals and use dedicated electronic feeds supplied by publishers. Scholar, by contrast, uses an automated process. Software robots crawl the web in search of documents that look like scientific papers, and then use algorithms to strip out relevant information such as author and publication date. The process is vastly cheaper and quicker, but it is not yet updated daily and there are no manual checks to delete duplicates or correct misclassified records.

Google has deals with several academic publishers that allow it to search the full text of many papers, whereas Web of Science and the others are largely restricted to searching abstracts. But Scholar's index is restricted to online sources — Web of Science has archives that go back to 1900. And the automated

process means Scholar's citation tracking can return odd results. For example, Web of Science finds almost 14,000 citations for a 1988 *Science* paper on the polymerase chain reaction[2], identifying it as the most highly cited paper ever to appear in that journal. Scholar finds just under 3,000.

All this suggests that there may be little overlap between the citations in the grey literature found by Scholar, and those extracted from the primary literature — even when the
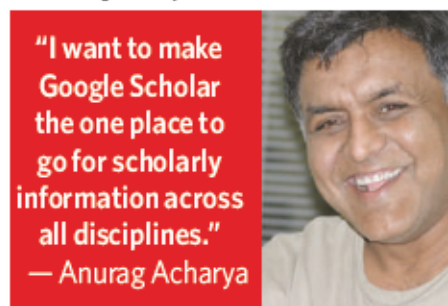
> "I want to make Google Scholar the one place to go for scholarly information across all disciplines."
> — Anurag Acharya

citation counts match up. For now, librarians are unanimous in their advice: stick to Web of Science or Scopus if you need to do a thorough literature search or an accurate citation count. The engines have impressive coverage and well indexed records with fewer misclassified entries. Librarians also warn that Scholar is still an experimental, or beta, version. Google remains reluctant to reveal details of its search algorithm, or what it indexes, so hopes of using Scholar as a tool for checking on citation counts is a distant prospect, they say.

All three search engines will continue to evolve. Scopus and Web of Science plan to add additional resources to their databases, such as institutional repositories, together with new ways for searching those sources. Scopus, for example, is integrated with a chemical database, such that users can go from a literature search to see structural information on molecules of interest. But it is unlikely that these engines will ever mine the grey literature as broadly as Scholar. Elsevier has a separate, free search engine, called Scirus, that searches science web

resources, but it doesn't track citations.

So where does this leave Acharya's bold goal? Librarians say that Scholar's current high usage rates are likely to reflect searches run by undergraduates, who typically require only a couple of key papers on any one subject, and researchers who want a quick snapshot of an unfamiliar field. Acharya says he intended Scholar to appeal to such users, but also wants to attract academics who need to keep up with the latest papers in their field. As Thomson and Elsevier continue to invest in new services, it will be interesting to see whether Scholar can keep up.

Two's company

With just two full-time staff working with Acharya, it would seem that Scholar is a low priority for Google. But maybe they could draw on the expertise of outside computer programmers by letting them write software that taps into Scholar's database. It is an approach Google has used before to good effect. If Google allows programmers to do the same with Scholar, it is likely that add-ons would be developed by librarians and academics.

So does Scholar plan to open itself to outsiders? Not right now, says Acharya. He remains cagey, but is not ruling it out. "We may reconsider this decision once the service is closer to how we envisage it."

The Google team may also reconsider if enthusiasm for Scholar continues to grow. Librarians at Virginia Tech in Blacksburg have already created a free software extension, called LibX, for an Internet browser, which allows users to retrieve papers using Scholar with a simple mouse click on highlighted text. LibX will take you directly to your library's resources, if the paper can be found there. And that is the sort of tool both Google and librarians can learn to love.
Jim Giles is a reporter for *Nature* in London.

1. Bauer, K. & Bakkalbasi, N. *D-Lib Magazine* 10.1045/september2005-bauer (2005).
2. Saiki, R. K. et al. *Science* 239, 487–491 (1988).

GOOGLE