Free standing
NIH director defends plans for open-access archive
p424

Game plan
Billion-dollar study intends to track 100,000 children
p425

Tissue issue
Biobanks urged to standardize sample accounts
p426

Fired up
Autistic arsonist convicted over pollution protest
p428

Science searches shift up a gear as Google starts Scholar engine

Declan Butler

Google has unveiled a test version of a search engine aimed specifically at academic material. First impressions of Google Scholar from librarians and computer experts suggest that the service is impressive in both scale and functionality.

"Google has made a very good start," says Ed Pentz, executive director of CrossRef, a collaboration of more than 1,000 publishers and societies that aims to improve online linking and access to papers.

The engine searches only research publications such as journal articles, books, preprints and technical reports, putting the most pertinent articles at the top of its searches by means of algorithms similar to those used by the firm's conventional web search. These analyse the number and importance of links pointing to sites (see *Nature* **405,** 112–115; 2000). In Google Scholar, papers with many citations are generally ranked highest, and they get a further boost if they are referenced by highly cited articles.

GOOGLE

Lord of the files: Anurag Acharya created Google Scholar to search academic papers.

A test drive of the beta engine, or test version, provides a peek under the hood. A conventional Google search for "human genome" throws back several million hits, with genome centres and databases ranked top. Google Scholar, by contrast, returns only around 100,000, with the landmark *Science* and *Nature* human-genome papers from 2001 both appearing in the top three.

Google says almost all "major publishers" have allowed the full text of their papers to be searched, although it declined to provide a list of those involved. The engine also searches abstracts from online archives such as PubMed and the NASA Astrophysics Data System, and the complete text of physics preprints on the arXiv server. In total, almost half a billion documents are thought to be covered.

The index is itself still highly incomplete, however. For technical reasons, large swathes of the often complex article databases supplied to Google by collaborating publishers are not covered. Google says it is working to solve these problems.

In addition, many papers can be only partially searched by the engine. Elsevier, the largest scientific publisher, has so far declined to allow Google to index its text, although the engine includes hits for more than a million Elsevier articles indexed as abstracts. "Google Scholar is an experiment, a beta version, and we are curious to see how it will be utilized and developed," says Marike Westra, a spokeswoman for the publisher.

Scope for competition

Elsevier is also marketing a commercial search engine, known as Scopus (see *Nature* **428,** 683; 2004); annual institutional subscriptions to this engine run from $25,000 to several hundred thousand dollars. The company argues that research institutions are willing to pay for high-quality search and information services such as Scopus and Web of Science, which is marketed by Thomson ISI of Philadelphia.

Some publishers may be concerned that Google Scholar has a subversive feature. Clicking on a hit returned by the engine takes the user to the article on the publisher's site. But Google Scholar also links to free versions of the article archived on other sites, such as authors' personal home pages. It is unclear how publishers will respond to this.

The creative force behind Google Scholar is Anurag Acharya, an Indian-born computer scientist who was on the faculty at the University of California, Santa Barbara, before he joined the company in 2000. Acharya first had to make his software identify and gather scientific papers from around the web using simple rules based on the common format of scientific papers, and then extract the title, abstract, authors and references.

Extracting references, which come in a variety of formats and are often full of mistakes, is key. Once references and papers are interlinked, it is relatively simple to apply algorithms to create indexes and rankings.

"I started the project because I wanted to build something better for researchers," says Acharya. Building automated citation indexes was new to him, but the scaling up was helped by his background in designing large-scale distributed computing systems. "Extracting information and references was the hard part," he says. "Building an index, making it run fast, and stable, that was easy; I already know how to do all that." ■

➧ http://scholar.google.com