Mining rich seams of information

From databases to desktop agents, there's more to search than Google. Belinda Weaver talks about tools and strategies.

Using Google for every Web search is like having nothing but band aids in your medicine cabinet. Finding information involves more than Googling. A specialist tool – a newgroups searcher, PDF finder or subject page – may be needed. Matching tool to task is one way to get results. Knowing what's out there is another.

Google doesn't index everything on a Web site. If information is stored inside a Web database, Google probably won't be able to peek inside. It can tell you the White Pages exist, but it can't look up a telephone number. Nor can it search the Australian Honours List.

Why not? Although tons of useful information is stored in thousands of different databases online, Google can't index it. This may be for several reasons. Most database information is served up 'on the fly' – generally in response to typed in queries – so is too transient to be indexed. Information protected by firewalls, fees or passwords is also 'invisible'. Search engines are also poor with non-text material, such as Flash animations.

The Invisible Web

This would not be so bad if databases were a negligible part of the Web. But the invisible Web is not only around five times the size of the visible; it also has some of the most useful stuff online. Bibliographic databases are searchable files of journal articles. They are a fantastic source of information on new research, and essential for academic study. Scirus (http://www.scirus.com/) is a giant gateway to article searches in the sciences. It provides links to high quality science sites as well.

Full-text databases include Stock Exchange Announcements, patents, industrial awards, transcripts, annual reports, and more. Product databases, such as bookseller Amazon's vast database, are searchable catalogues of goods or services. Scientific databases include lists of animal species, plants, and the periodic table of the elements. The four-yearly census is a statistical database. Aviation crash statistics are another.

The inability of even the biggest search engines to mine this 'invisible Web' is their biggest limitation. Google, AltaVista, Teoma, AllTheWeb – they make the visible Web navigable, but miss many resources of substance. To find them, use tools such as Gary Price's Direct Search (http://www.freepint.com/gary/direct.htm). Though not very user-friendly, DirectSearch shows the scale of what the search engines miss. Price has written a book, *The Invisible Web*, co-authored by Chris Sherman, who writes the SearchDay (http://www.searchenginewatch/com/ searchday/) newsletter. The book has a companion site (http://www.invisible-web.net/). Profusion (http://www.profusion.com/) is another iWeb tool.

Google is trying to solve this. CEO Eric Schmidt said in 2002 that he wanted Google to index material locked up in premium databases like Lexis-Nexis. While Google has since added PDF search and searches of Word documents, this 'gateway to everything' dream is still some way off.

Anyway, the answer to successful web searching is not so much a better engine, but a strategy that involves different tools and probably a different mindset.

Whenever you search, think first about what you hope to achieve. Are you looking for a fact? A brief introduction to a subject? Deep background? You will need different strategies for each.

In the first two cases, an online encyclopaedia will be more useful than Web searching. Encyclopaedias structure information into manageable chunks, and cover an enormous range of subjects. Topic-specific ones such as the PC Webopedia (http://www.pcwebopedia.com/) can also be handy.

Virtual reference desks

To find look-up tools, use Virtual Reference Collections. VRCs contain everything you'd expect a library to have – dictionaries, thesauri, encyclopaedias, maps, atlases, factbooks, biographical dictionaries – as well as currency converters and translation tools. Go here for basic, factual information. It's quicker than searching and the information is more reliable. A site such as the CIA's World Factbook will answer an enormous range of questions – all in 2003 figures.

Try the Internet Public Library's Ready Reference collection (http://www.ipl.org/). Direct Search includes ready reference. My own Guide to Internet information sources for Australian journalists (http://www.journoz.com/) provides basic tools and a gateway to Australian sources of information.

Defining terms

The right search words can make all the difference. With any kind of medical question, always look for both the common and clinical names, e.g. 'shingles and 'herpes zoster'. Master the use of the Boolean operators AND, OR and NOT, as these provide a structured way to link terms effectively. AND allows you to narrow down searches. If you search for 'economy AND Australia', the results will include only sites where both terms appear. Google (http://www.google.com/) does not offer true Boolean but does operate with an implicit Boolean AND. The NOT operator excludes terms, and must be used with caution.

Using OR broadens a search. Use it when you are not sure what terms will be successful, e.g. 'asylum seekers OR refugees OR illegal immigrants OR boat people'. Once you have got some results, try to add new terms to further narrow things down. Clustering engines can be good for that, as they suggest new concepts. On new tool Mooter (http://www.mooter.com/), a search for 'music downloads' gave clusters for MP3, software, bands, sheet music, lyrics and more. Vivisimo, another clustering tool, can also give you what's top in several engine databases (http://vivisimo.com).

In Advanced Search, HotBot (http://hotbot.com/), which searches Google, AskJeeves, Lycos and its own database, and AllTheWeb (http://alltheweb.com/) simplify Boolean by offering drop-down options. 'All of the words' equals the Boolean AND search; 'any of the words', the Boolean OR. Exact phrase' (these words in this precise order) is useful for proper names or concepts always expressed as a phrase, such as 'open source programming'.

Limiting search results to PDF can be a way of weeding out flaky material. Material produced in PDF tends to resemble traditional publishing, so PDF search can be a quick way to find reports, for instance.

Information strategy

Though the right terms make a difference, try to think exactly *what* you need to know. Then try to identify what kind of organisation could tell you. Information doesn't come out of a vacuum. It is generally produced so people can tell you something, or sell you something. Finding organisations, which have names, goals and mission statements, is easier than searching for abstract 'information'.

Want the facts on immigration? If you think 'Who would be gathering this kind of data?', it's not a huge leap to the answer – a statistical agency. Australia's agency is the Australian Bureau of Statistics. They have exhaustive tables of information not just on immigration, but about every aspect of Australian life. Much of the data can be compared over time. Many state and university libraries provide free access to ABS publications.

When the SARS scare was at its height, information was everywhere. But bodies such as the World Health Organization (http://www.who.int/) had the latest, most authoritative information. Finding the kind of organisation that will be publishing the information you seek really pays off. Use the directories at http://www.journoz.com/orgs.html to identify them.

Little helpers

Improve your search speed and access to look up-tools with browser and desktop helpers. Google's browser toolbar allows you to run Web searches from anywhere. AltaVista's toolbar (http://www.altavista.com/toolbar/default) offers translation as well as multiple kinds of search – for video, audio, images, news and searches of Australian content.

HotBot offers a multitasking DeskBar that incorporates dictionary, encyclopaedia and quotation lookup as well as Web search. Google's new deskbar can run searches and lookups while you are working in other applications. Like HotBot's, it sits on your Windows taskbar, so is 'always on'. Quick and easy to install, these tools earn their keep straightaway. Keep tabs on what Google is up to at Google Labs (http://labs.google.com/).

Other desktop tools include searchbots such as WebFerret (http://www.webferret.com/) and CopernicAgent (http://www.copernic.com/). These can be customised, for example, by saving search strategies for reuse or querying certain types of service. With Copernic Agent, it's possible to store multiple search criteria. Searchbots also cover more of the Web, and remove duplicate search results. Some even mine invisible Web resources. Sites like AgentLand (http://www.agentland.com/) and BotSpot (http://botspot.com/) cover customisable 'bots', including shopping bots and personal assistants.

Finding news

Sometimes the background you need will have been covered in magazines or newspapers. Dedicated tools such as NewsText (http://www.newstext.com.au/) or the NewsStore (http://www.newsstore.com.au/) provide a way in. Thought not free, you may be able to access these services through subscribing libraries. Free archives such as the

UK Guardian (http://www.guardian co.uk/) or the multimedia archives at BBC News (http://news.bbc.co.uk/), provide background on world events. Free magazine articles can be found at FindArticles.com (http://www.findarticles.com/), which provides access to more than 700 magazines, including *Macweek* and *Wireless Review*.

Expert helpers

Another search shortcut is to use subject pages rather than search. Yahoo! is probably the best known subject page, but topic-specific ones such as Austlit (http://www.austlit.edu.au/) or EdNA - Education Network Australia (http://www.edna.edu.au/) will steer you to selected sources, saving endless sifting and sorting. A page of links compiled by an expert is gold. Use the Librarians' Index to the Internet (http://lii.org) or Pinakes (http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html).

AskNow (http://www.asknow.gov.au/) is another useful tool. Funded by government, AskNow exists to answer *your* questions – for free. It's online ten hours a day, it works in real time via chat, and you can ask anything. A trained librarian will help you and will email you a transcript at the end. Google Answers (http://answers.google.com/answers/) offers fast Q&A too, but you have to pay.

Staying ahead

Monitor sites such as SearchEngineWatch (http://searchenginewatch.com/) to stay on top of new developments. The site has Web searching tips, and a comparative chart (http://searchenginewatch.com/facts/article.php/2155981) of search engine features, such as the ability to use truncation and wild cards, or options such as 'Find similar' or 'Search within results'. Use the power searching tips to beef up your skills Search engine charts and features are also on offer at Search Engine Showdown (http://searchengineshowdown.com/).

Finally, use a checklist to ensure you have looked as far as you can. There are four key areas for publishing – government, organisations, the educational sector (new research) and business. Yahoo! (http://www.yahoo.com.au/) is the best tool for .com links. Use EdNA (http://www.edna.edu.au/) for .edu. Governments on the WWW (http://www.gksoft.com/govt/en) will deliver .gov. Directories (http://journoz.com/orgs.html) will link you to organisations (.org and .asn). It is always quicker to search a site that has information you want than to search more generally and waste time sorting good links from bad.