

Bench philosophy (19): Literature databases

# The Art of Literature Search

Searching for scientific literature has never been easier. Journals are keener at placing their content online, and search engines are getting smarter and more powerful. Sophisticated indexing and even natural language processing, mean that finding what you are looking for is only a mouse click away. But are you keeping up with it all?

Let us start with the obvious. We are in the Google age and biologists are just like anyone else. If you want to know about a topic, just Google it! Resist the temptation to be snobbish about Google Scholar (GS). GS (<http://scholar.google.co.uk/>) uses the Google page ranking system, so it takes you straight to the key papers. On top of that, each hit is actually a combination of different versions of the same citation, making it easier to find what you are looking for. It also gives you the option of limiting the findings to recent citations, and the key authors in the field are listed at the foot of the page for follow up. Because it is built upon the Google search engine, it doesn't place a limit on the number of search terms, the language of the sources or (provided the document is at least cited online) the extent of journal coverage.

Google Scholar has other uses too. Because it covers the whole web rather than just literature repositories, it is also a good place to try if your preferred database doesn't have a link to the full text article. After all, the authors may have posted a copy of the paper on their website. But beware: that version may not be exactly the same as the official published copy. GS also figures out what full content you are entitled to. If you access GS from an institutional address, it works out which subscriptions your library has access to and automatically inserts a link to the full content next to the search hits.

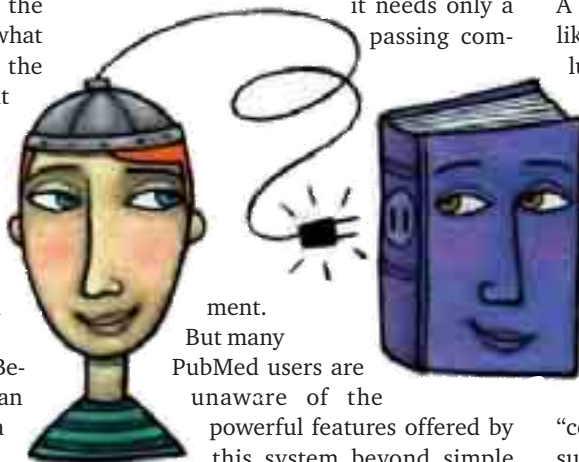
## Filtering out information

GS isn't the only web server dedicated to science. Scirus (<http://www.scirus.com/>) is another web-based search engine dedicated to scientific literature and, therefore, like GS, provides access not only to published literature in peer-reviewed journals but also to a host of other publication types, such as course material, scientists' web pages, pre-prints, patents and institutional repository and website information. The results can be filtered by sources, while journal sources can be filtered separately from websites.

You can also filter by file type (pdf, html or word).

For serious literature research, though, most biologists would probably head straight for PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>), the freely available literature database offered by the National Library of Medicine. This resource, the oldest of its kind, is so familiar you may think

it needs only a passing com-



ment. But many PubMed users are unaware of the powerful features offered by this system beyond simple keyword or author searches.

For instance, take a look at that box labelled "Search" at the top on the left. After you have done a basic text search, try following it up by using one of the options in this dropdown box. An OMIM (Online Mendelian Inheritance in Man) query may show you some surprising diseases you didn't know your search term is related to. Likewise, selecting the Gene option may throw up some surprising genes related to your subject. Alternatively, why not do all the options at once by selecting the "All Databases" button, top left. This effectively widens your search beyond just literature to the whole of the bioinformatics field. After all, the published literature is just a tiny part of the broader bioinformatics knowledge base. This integrated bioinformatics approach also underlies the EBI (European Bioinformatics Institute) server at <http://www.ebi.ac.uk/>, the European answer to PubMed.

Surprisingly, one of PubMed's major strengths is also one of its most underused

features. People forget that it is an indexed database: human beings have gone through the articles and categorised them according to their subject. These subject categories are called MeSH (Medical Subject Headings), and they allow you to search by idea, rather than by word. For example, imagine you want to find out about the effects of alcohol on development of the cerebellum. A text search string might be something like "effect alcohol development cerebellum". However, a paper that used the word "ethanol" instead of "alcohol", would be missed. The MeSH system can be used to map words to concepts in a guided manner. To pursue our cerebellum illustration, click on the "MeSH database" on the blue bar on the left of the PubMed main page and enter "alcohol" in the search box. The top two MeSH headings are "ethanol" and "alcohols". Select those and add them to your search box, then do the same for "cerebellum" and "development". The result is a much more focussed list of references and can be crafted as precisely or as broadly as you want, depending on which MeSH terms best matched your interest.

## Database suits

PubMed and EBI can best be thought of as bioinformatics-based access to biological research. They are really a suite of databases that work together to deliver integrated information all the way from gene through protein, to whole animal. The literature is really just a part of this grand whole. Scopus (<http://www.scopus.com/home.url>) and Web of Science (<http://apps.isiknowledge.com/>), on the other hand, are databases that focus especially on the literature. As such, they have enhanced literature-based features, such as citation analysis, which allow you to discover which references cited a particular paper, or perhaps find related papers based on shared citations. Another way of expanding a literature search with Scopus is by selecting a citation's keywords and finding references that share the same keywords. In other words, with these data-

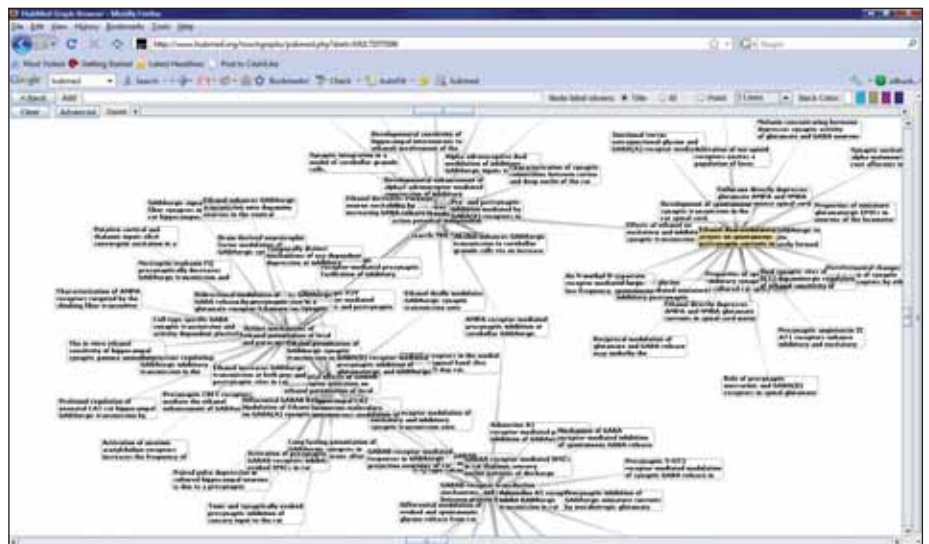
bases it is possible not only to do a search but also to analyse the corpus of citations. For instance, with Scopus it is easy to get a profile of which journals publish in which areas. You can also look up an author and get a profile of how many times their papers were cited year by year. In addition, Scopus also returns scientifically-relevant web results (powered by Scirus) and patents through a set of parallel tabs.

We have already seen the merits of enhancing a simple text search with a subject-heading search in PubMed. But what about doing the same thing with Gene Ontology (GO)? That is the special feature of the GoPubMed server (<http://www.gopubmed.com/>). GoPubMed takes your query and not only returns the PubMed hits but organises them according to their GO entries. This makes the process of drilling down to the subject of interest simply a matter of selecting terms in the GO tree for inclusion or exclusion and the server updates the search query automatically. On the same website there is also GoWeb, which does the same thing for the whole web, as well as GoGene, although GoGene is for subscribers only. A new database, Novoseek (<http://www.novoseek.com/Welcome.action>), works in a similar way, comparing your search term with its biomedical dictionaries and taking synonyms into account. Like GoPubMed, it also allows you to narrow your search by concept.

PubMed, GoPubMed and Novoseek organise search hits according to controlled vocabularies. But what about letting the search hits organise themselves? Clustermed (<http://demos.vivisimo.com/vivisimo/cgi-bin/query-meta?v%3aproject=clustermed&&v:form=frontpage=1>) uses powerful natural clustering technology to organise hits according to phrases in the abstract. The clusters are displayed hierarchically and clicking on any node in the tree takes you to the abstracts with links to the PubMed sources. It also works in reverse: each abstract also has a link to show you to which other cluster nodes the abstract was allocated. Even looking at the cluster tree without following any of the abstracts is, in itself, a simple way to get an overview of a topic – the headings tree for your next review article!

### Hubmed instead of PubMed

A blogger described Hubmed (<http://www.hubmed.org/>), another enhanced literature database, as “PubMed on Steroids” ([http://efficientacademic.wordpress.com/2006/01/30/hubmed-PubMed-on-](http://efficientacademic.wordpress.com/2006/01/30/hubmed-PubMed-on-steroids/)



**Hubmed-touchgraph-complex:** Clicking on nodes in the hubmed touchgraph applet can result in a network of related articles.

*steroids*), although the site describes itself more modestly as “an alternative interface to the PubMed medical literature database”. At first, it does indeed look like a stripped down version of PubMed. But take a closer look and the reason for the blogger’s enthusiasm gradually becomes clear. For one thing, with Hubmed you can tag articles. Click on the “tags” button at the top right and you are taken to a community-built set of tagged articles, or sign in and apply your own tags. Exporting references to your reference manager is even easier than PubMed: just click on “export” and the server does the rest. Or perhaps you have been reading a pdf and come across an interesting citation in the references section. How do you get hold of the original? Copy and paste it into Google Scholar? Better still, just paste it into Hubmed’s “Citation Finder” and it takes you straight to the reference.

One of the most enjoyable features of Hubmed is the touchgraph: a visual representation of related articles arranged as a graph. Tick on the box of an article you are interested in, scroll down to the bottom of the page, and hit the “Touchgraph” button. Give Java a few seconds to load the applet and you will see your abstract come up as a node on the middle of the page. Double click the node and it will expand into all the related abstracts. To read any of these abstracts, simply click on the red “info” button at the top of each one. As you carry on clicking on the nodes, the network goes on expanding. Admittedly, things can get a bit noisy after a while, and there will be a lot of nodes you are not interested in but you can hide nodes you don’t want with a simple right mouse click. The end result is

something like a concept map. Some databases straddle the borderline between being a database and a data-mining system. Xplormed (<http://www.ogic.ca/projects/xplormed/>) is a powerful example of what can be achieved when humans and machines work together, using the relationships between words and ideas to produce a focussed reading list. Xplormed begins with a MeSH-based search but the difference is that the hit-list is analysed for strings of “word chains”. From a list of these word chains, you select the ones that most closely match your interests. For example, if you were interested in finding out the effects of alcohol on cerebellar purkinje cells, the word chains “purkinje, cell, loss” would be worth following. The link brings you to a list of highly relevant papers that can, in turn, be narrowed down and the whole process reiterated until you get the list you want.

The web is an exciting place of experimentation. New, widely different ways of searching are being developed all the time. But in the face of all this variety and innovation, one simple point stands out: simple text searching by keyword is not enough. To keep on top of the information explosion means deploying a range of search techniques and always being open to new ways of searching.

STEVEN D. BUCKINGHAM

**Fancy composing an installment of “Bench Philosophy”?**

Contact Lab Times  
E-mail: [editors@lab-times.org](mailto:editors@lab-times.org)