EUR J PHYS REHABIL MED 2008;44:221-7

To Google or not to Google, this is the question

E. GIGLIA

University of Turin, Turin, Italy

To Google or not to Google has become in the latest years a big question for researchers: immediate answers, but often thousands; sometimes pertinent, sometimes not reliable or too commercial-oriented. No doubt that Google always "finds" something — and sometimes it is the only way, or it retrieves real pearls — but where does it search? How can a researcher refine or limit the search? That is why we'll explore some Google features not so widely known, and other search engines with useful devices to perform a more efficient search in the biomedical field.

KEY WORDS: Internet - Google - Research.

Last time we talked about subject gateways and metasearch engines crawling among resources already selected and evaluated by information specialists. Now we're going to have a look on search engines, generalist which hunt for any information all over the Web – or specialized, dedicated to scientific resources.

Please, raise one's hand who doesn't begin any search on the Web googling a keyword. "To google" stood out as a neologism in English, due to the popularity the search engine reached.[1] In the British 2005 survey *Open access self-archiving: an author study* (Key Perspectives, http://eprints.ecs.soton.ac.uk/10999/1/jisc2.pdf), 72% of the interviewed academic researchers declared that they start from Google to search the Web also for scholarly articles. "Googling" has deeply influenced the users' expectations about the answers to their questions and the way of searching the Web – that means that it has, or ought to have, deeply influenced also the librarians' job, but this is another matter… Needless to say we can't ignore these premises, so we shall start from Google itself and its advanced options. Aim of this contribution is to show that Google is not alone, and that other search engines could be useful for scholarly researches, both because they are dedicated to explore only biomedical resources, or because they offer features — like clustering the results — that make life easier. Don't forget that there is no recipe at all: the search depends on the question, the aim and the perspective of the query. And keep in mind that search engines perform a generic search in the Web, retrieving composite resources such as web sites, free articles, and commercial materials: there are no filters of quality or of reliability, that's why the question to Google or not to Google has taken its relevance in the biomedical field.

Address reprint requests to: E. Giglia, University of Turin, via Sant'Ottavio 20, 10126 Turin, Italy. E-mail: elena.giglia@unito.it

Advanced Google: limits and filters

We're going to see an "unpublished" Google: the Advanced Search mask (Figure 1), and how it allows you to refine your query. First of all you can associ-

Figure 1.—Google advanced search.

ate words with the Boolean operator, *i.e.* you can tell Google to find ALL the words you type (entering AND), *e.g.* stroke AND rehabilitation, 433 000 results; AT LEAST ONE of the words (entering OR), *e.g.* stroke OR rehabilitation, 10 800 000 results; WITHOUT a word (entering NOT). If you want you can also search for the EXACT PHRASE – in this case you can get the same results from the traditional Google homepage search box including your phrase in brackets, *e.g.* "stroke rehabilitation": 245 000 results.

You can also limit your search by:

— format: if you choose the .pdf one, you will be easier returned journal articles freely accessible on the Web: *e.g.* "stroke rehabilitation" in .pdf limits to 104 000 results. Other choice is the .ppt format, giving presentations in congresses or meetings or academic lessons;

— domain: *e.g.* if you limit to the .edu domain, only academic URLs will be considered, or if you exclude the .com domain you will not be shown commercial sites;

— occurrences: depen-ding on the purpose of the query, you can ask Google to search your keyword anywhere in the page, just in the title, or in the text, or in the URL: *e.g.* stroke rehabilitation just in the title and in the .pdf format limits to 710 results (from 433 000 of the simple search);

— date: be aware that "date" stands for the time the page was first seen by the search engine's spider, not for the date of creation or update!;

— usage rights: you can ask to see only material free to use or share (according to the terms of the associated Creative Commons licences), that is useful if you don't have an affiliation to an institution that provides you with online subscriptions to scholarly resources.

Of course, you can apply one or more limits to the same search, in order to obtain the narrowest result, without noise.

If you need a definition, Google can provide for it, if you type the operator "define:" with no spaces between it and the term you want to be defined, it will show you a list of definitions gathered from various online sources: *e.g.* "define: scoliosis" get a list of more than 20 different definitions from medical glossaries free on Web.

In the Web 2.0 momentum, Google is also useful to find out blogs or newsgroups which are not to be underestimated about subjects of interest. The right URLs to start from are, respectively, - http://blogsearch.google.com/ and - http://groups.google.com.

Google Scholar: standing on the shoulders of giants, but…

Google Scholar (http://scholar.google.com/) is a service provided by Google aimed at finding peer-reviewed papers, theses, books, abstracts and articles from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. The interface is the usual easy search box of Google, the ranking of the result is quite the same of the general search engine: weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature. But… Does it work applied to scholarly literature? Perhaps if I'm looking for the most cited article, it does, but if I'm looking for the most recent one, the criterion of the most cited one doesn't work at all, because it would retrieve only old stuff. It is true that in the Advanced Search page you can limit by date, asking for articles published between a range of years, and that from the results page you have the option "Recent articles" in the green toolbar on the top, but… as you can see in

Figure 2.—Google Scholar – results for "spastic hemiplegia".

Figure 2, results for "spastic hemiplegia" list articles written in 1987, 2001, 1991, 1988, 1993, and so on, while the "recent" ones gives you 2003, 2006, 2005, 2003, 2002 items.

This emphasizes another evident lack of Google Scholar: it never indicates its coverage, nor the period of embargo that the publishers establish, *i.e.* you never know what you are searching, and where, since when and till when. So, you stand on the shoulders of the giants, as the idiom states, but you don't know neither who the giants are nor their age or weight…

What Google Scholar provides in an excellent way is the Citation Tracking feature: for each item it shows the link "Cited by" that allows tracking back the history and the relevance of a work. A survey conducted by Maurella Della Seta and Rosaria Cammarano, of the Italian Istituto Superiore di Sanità on *Citation tracking of scientific publications through two different searching tools: Google scholar and Web of science.* (http://eprints.rclis.org/archive/00008271/), showed that Scholar is a valid complementary tool of the very expensive Web of Science database. Resulting citations do not match exactly, due to the difference in type of documentation considered by search algorithms: the overlapping is almost of the 50% and in the years Scholar had enhanced the number of unique citations retrieved.

A last positive annotation about the coverage of Google Scholar: being a free-of charge service, it harvests the Open Access repositories, so it lists with a good relevance ranking scholarly material self-archived in full-text by the authors (*i.e.* the pre-print of the final versions submitted to the traditional journals).

Google Co-op: a new, social tagging Google way to explore the Web

Google Co-op is a project – in Beta phase yet – within which a researcher can use his/her expertise to improve Google search. There are some "Topics", among which "Health" (http://www.google.com/ coop/topics/Health), that are specific search areas that Google is developing with the help of expert contributors. Contributors to topics annotate websites that they think are especially useful, relevant, or authoritative: the labels appear as links at the top of search results pages when users search for something related to the topic. Users can click these labels to refine their search

Figure 3.—Google Co-op Health home page.

results. Possible choices in "Health": treatment, prevention, practice guidelines, test/diagnosis, and so on (Figure 3). The project combines the easiness and speedy of the Google search with a sort of quality filter applied by specialist. A good tool to keep under observation.

OAIster and Scientific Commons: Open Access search engines

OAIster

OAIster (http://www.oaister.org/), powered by the University of Michigan, is the first search engine dedicated to Open Access resources (for Open Access policies see Appendix I). It provides access to about 15 million records from more than 900 repositories of scholarly institutions. Digital material searched by OAIster include also thesis and doctoral thesis, datasets, digitized books and articles, videos, images, audio files, etc. Filters are possible by title, author, subject, language.

According to the principles of the Open Access movement, all of the contributions ought to be freely available in full text, even if sometimes in a pre-print version, following the copyright permission of the different publishers. The great utility of this engine is that you can find the pre-print of an article otherwise "closed" in a subscription-fee journal. For instance, if you run a search for "low back pain", limited to the title, and to text as resource type, you will find 771 items, among which the article *Frequency of low back pain among men and women aged 30 to 64 years in France. Results of two national surveys*, only accessible to the subscribers of the «Annales de réadaptation et de médecine physique». Or, for example, you can find a technical report not published elsewhere: *Image Registration and Statistical Analysis for Quantitative In Vivo Spin-lock Magnetic Resonance Imaging of the Intervertebral Disc Response to Compression*, submitted by a staff of the University of California, Berkeley.

Scientific Commons

Scientific Commons (http://www.scientificcommons.org/) is a project of the University of St. Gallen (CH). This is also a search engine dedicated to Open Access resources, and to indexes about 16 million records from about 900 repositories. Aim of the project is to develop the world's largest communication medium for scientific knowledge products which is freely accessible to the public. The search by keyword is very effective, and presents filters by date or language.

Biomedical search engines

MedHunt

MedHunt (http://www.hon.ch/ MedHunt/), powered by the Health on the Net Foundation (see the last number of this column), is a search engine dedicated to biomedical resources. It deals with web sites, web pages, and also some free articles. The great advantage is that some of these resources are evaluated by the HON staff of information specialists.

OReFiL

OReFiL, Online Resource Finder for Lifesciences (http://orefil.dbcls.jp/index.cgi), developed at the University of Tokyo, returns up-to-date query-relevant online resources introduced in peer-reviewed papers, extracting URL from Medline abstracts. You can run a query by free words, MeSH terms or authors' name and you will find datasets, relevant web pages, and related articles.

TRiP

TRiP, Turning Research into Practice (http://www.tripdatabase.com/index.html), is a search engine specialized in evidence-based resources. It clusters the results in categories like Systematic reviews, Guidelines, and so on, and, for Medline articles it offers also a clusterization by Subheading (Therapy, Diagnosis, etc.). A filter by specialization is also allowed, to obtain a cluster for the "Specialist primary research" in the requested medical specialization, *e.g.* ortho-paedics. A nice new feature is that by clicking on the "i" icon you can view the article's Conclusion, a new useful tool instead of the common Abstract view
(Figure 4).

Figure 4.—TRiP search for "hip fracture", clustered by systematic reviews, synopses, guidelines, etc.

Clusty and Biometacluster: a generalist search engine and its biomedical spin-off

Clusty

Clusty (http://clusty.com), powered by Vivísimo, is a generalist metasearch that applies the cluster technology to split items in an effective way. You can use it for your everyday query, viewing not only your results clustered, but also the domains of the sites ("Sites" label on the right), which indicates how many results you have for *e.g.* for the .edu material.

BioMetaCluster

BioMetaCluster (http://vivisimo.com/html/ biometacluster), is the biomedical project linked to Clusty. It explores several sources — you can check and flag them from the Advanced Search — like Wikipedia, PubMed, OMIM, the site of the FDA (Food and Drug Adminis-tration, USA), Clinical trials.gov, ACS (American Chemical Society), and news sites like CNN, Reuters and so on. As you can see in Figure 5, for a search on "spinal cord lesions" BioMetaCluster splits the results among metastases, multiple sclerosis, posting, and so on, harnessing the cluster technology to help the user to orient between the retrieved items.

Metasearch: seeking after the biomedical world

Scirus

Scirus (http://www.scirus. com/ srsapp/), searching more than 450 million science-related pages, focuses only on web pages containing scientific content and goes deeper than the common search engines, looking for reports, peer-reviewed articles, patents, pre prints and journals. It allows also to select a subject area, to narrow the searching by author, journal, date, and to customize and save searches.

The Diseases database

The Diseases database (http://www.diseasesdatabase.com/begin.asp) is a cross-referenced index of

Figure 5.—BioMetaCluster and the results for "spinal cord lesions".

human disease, medications, symptoms, signs, and abnormal investigation findings. It gives definitions, links to special web sites, or sends your search to other resources like Wikipedia or Scirus itself.

Googling PubMed: new interfaces to ask a myth

As we said, Google has deeply changed the users' expectations in terms of easiness, speedy and efficiency of the query. Assuming that, and harnessing the potentiality of the Web 2.0 and the semantic Web, some new projects offer a Google-way to search PubMed.

GoPubMed

GoPubMed (http://www.gopubmed.org/), is far the most interesting experiment in providing a new, inviting interface to navigate PubMed and its lists of results. GoPubMed is a knowledge-based search engine for biomedical texts. The *Gene Ontology*, a controlled vocabulary to describe gene and gene product attributes in any organism, and the MeSH controlled vocabulary serve as "Table of contents" in order to structure the 16 million articles of the Medline data base.

GoPubMed retrieves PubMed abstracts for your search query in the same way that PubMed does, but then sorts relevant information to the four top level categories:

— WHAT: detects terms from the Gene Ontology (GO) and Medical Subject Headings (MeSH) in the abstracts, displays the subset of the GO and MeSH relevant to the keywords, and allows you to browse the ontologies and display only papers containing specific GO and MeSH terms. By navigating the tree you can narrow down from thousands of search results to a few in seconds, reducing your waste of time;

— WHO: this category helps you to find scientists and centres;

— WHERE: you find geographic localization of persons, centres, universities. The journals for the query are also listed in this category;

— WHEN: "is the citations time machine", can change the time window for the search in a while.

In order to improve the performance of the text mining, each researcher is invited to become a "curator" in the "Folksonomy 4 Science" project: you can register and you will be asked to mark highly relevant terms as well as terms which have a differing meaning in the abstract of articles.

Figure 6.—GoPubMed and its new way to navigate PubMed results.

Navigating the tree of the extracted concepts you can get an answer to your question, *e.g.* "Which are treatment outcomes for the spine traumatic injuries?". PubMed retrieves more than 2 000 items for "spine traumatic injuries", GoPubMed then allows you to go further and intuitively refine your search just by clicking on the corresponding label in the navigation tree on the left (without using the Details box like in PubMed), till you get the answer you were looking for, as shown in Figure 6. Please notice that the system shows directly in the first places pertinent items which in PubMed run 72, 106, 122… clearly over the point you actually read the results list.

Hubmed

Hubmed (http://www.hubmed.org) is an alternative way to ask PubMed. It has a simple Google-like search box, but offers the possibility of clustering the results. It can also associate to each article a tag, different from the MeSH, assigned by the users, in the logic of the Web 2.0.

ReleMed

ReleMed (http://www.relemed.com/) is another search engine that runs on PubMed, using a different retrieval system, which promises you to show you the same results of PubMed, but with a better relevance ranking gathered from your query.

eTBLAST

Finally, eTBLAST (http://invention.swmed.edu/etblast/) a unique search engine that doesn't work with keyword but with phrases: it lets you input a whole paragraph and returns abstracts in PubMed that are similar to it. A curiosity: this search engine has been used to rate the similarity between scientific articles, with the surprising outcome that in too many cases they look like too equal, or better… plagiarized? You can read the very instructive *A tale of two citations* in Nature, 24 January 2008…and, maybe, rethink about the present scholarly communication.[2]

References

1. *To "Google it" is to search the Internet*, CNN Money, February 1, 2008. Available from http://money.cnn.com/news/newsfeeds/articles/newstex/IBD-0001-22724253.htm
2. Errami M, Garner H. *A tale of two citations*. Nature 2008;451:397-9. Available from: http://www.nature.com/nature/journal/v451/n7177/full/451397a.html

APPENDIX I.
News from the Open Access world

US National Institutes of Health, European Research Council, Italian Istituto Superiore di Sanità, Harvard University mandates Open Access for their funded researches

- January 2008 was a profitable month for Open Access policies for several institutions, mostly in the biomedical field.
  - — Jan. 11: the US National Institutes of Health (NIH) mandated at last OA for NIH-funded researches. It applies to all peer-reviewed articles that arise, in whole or in part, from direct costs funded by NIH that are accepted for publication on or after April 7, 2008, i.e. it applies also to previous grants that are still generating new articles. The deputy deposit is of course PubMed Central, the digital archive of the NIH. The policy is presented in http://grants.nih.gov/ grants/guide/notice-files/NOT-OD-08-033.html
  - — Jan. 11: the European Research Council (ERC) released its mandatory OA policy. ERC requires that all peer-reviewed publications from ERC-funded researches be deposited into an appropriate repository – both institutional or disciplinary, such as PubMedCentral – and subsequently made OA within 6 months of publications; the same applies to primary data – *e.g.* nucleotide/protein sequences, anonymized epidemiological data, etc. to be deposited in relevant databases preferably immediately, never later than 6 months after their publication. The policy is available at http://erc.europa.eu/pdf/ScC_ Guidelines_Open_Access_revised_Dec07_FINAL.pdf
  - — Jan. 17: in Italy the Istituto Superiore di Sanità (ISS, National Institute of Health) adopted an OA mandate, requiring its staff researchers to deposit their peer-reviewed manuscripts in the Institutional Repository, for OA release with 6-24 months embargo: more in https://mx2.arl.org/Lists/SPARC-OAForum/Message/4178.html
  - — Peter Suber's SPARC Open Access Newsletter of February, 2008 deals with all these policies, and others like the Howard Hughes Medical Institute, the Canadian Institutes of Health Research: read more in http://www.earlham.edu/~peters/fos/newsletter/02-02-08.htm
- On February, 12th the Harvard Faculty of Art and Science adopted a OA mandate for all the scholarly articles of the members of the Faculty. But Harvard goes further, because it mandates also the copyright retention: read more on point 5, http://www.fas.harvard.edu/~secfas/February_2008_Agenda.pdf .