Information Retrieval: Literature Searching in Today's Information Landscape

Bianca M.R. Kramer*

In this era of exponential growth in the amount of scientific information, the ability to perform efficient literature searches is a key skill for students and researchers. This review covers essential aspects of literature searching, including devising a search strategy and selecting sources to search. Examples will be drawn from the field of biomedical research in general and evidence-based medicine in particular. The implications of the semantic web, Web2.0 and Open Access on literature searching will be discussed, and the role of academic libraries in both the dissemination of scientific literature and information literacy education will be highlighted.

*Citation: Kramer B.M.R. Information Retrieval: Literature Searching in Today's Information Landscape. *Hypothesis* 2010, **8**(1): e6.

Introduction

Today's information landscape differs strikingly from that of a generation ago. In the biomedical sciences, PubMed (1) (which offers access to the database MEDLINE) and other search engines provide instant access to millions of citations and abstracts of the world's biomedical literature. Information on any subject, ranging from fundamental research on cell cultures and lab animals to stage III clinical trials and systematic reviews, is readily available. This abundance of scientific information requires that students and researchers acquire key skills in information literacy, including the ability to select key words from a search question and to quickly assess search results for validity and relevance.

Efficient Search Strategies

In the first half of April 2010, on average 2475 articles were added to PubMed daily (2). Clearly, searching this constantly growing body of scientific literature in an efficient way is more important than ever. In literature searching, the two main goals are: (a) to retrieve only those papers that are relevant to the search question (this is called "specificity"); and (b) not to miss any of those papers (this is called "sensitivity"). These goals are especially important in conducting the rigorous searches that are required in evidence-based medicine (EBM). EBM is an approach to medical research and practice that aims to support clinical decisions by the best available evidence from systematic research that has been conducted using sound methodologies (3,4).

In devising a search strategy that is both sensitive and specific, it is helpful to identify

*Information Specialist Health and Medical Sciences, Utrecht University Library, Utrecht, The Netherlands.
Correspondence: **b.m.r.kramer@uu.nl**
Received: 2010/05/02; Accepted: 2010/06/17;
Posted online: 2010/07/30

the different components of the question at hand, and formulate as many search terms (including synonyms) for each component as possible. A separate search is then performed for each component, combining all synonyms with the Boolean operator OR. These steps, if executed properly, enhance sensitivity. Subsequently, the separate searches are combined using the Boolean operator AND. This step ensures that only those references that match all components of the search question are retrieved, thus enhancing specificity.

In EBM, search questions are often formulated according to a format called PICO (5): Patient, Intervention, Comparison, Outcome. A more generally applicable format is the three-part question or DDO (6,7): Domain, Determinant, Outcome (see Table 1) (6,7). In identifying search terms and synonyms for each component, it is useful to consider cross-cultural differences in nomenclature and spelling, trade names, and nomenclature that has changed over time.

Many search engines offer extensive thesauri to facilitate searching. In employing such thesauri, such as the MeSH (Medical Subject Headings) database (8) for use in MEDLINE and PubMed, a search is automatically broadened to include synonyms and word variants of the term that is searched. As citations are usually indexed (awarded thesaurus terms) based on the full-text of the article, a search that includes use of the thesaurus could yield citations that might otherwise have been missed. This is especially relevant in cases where no abstract is included in the database itself.

However, there are also significant drawbacks to using thesauri (9). In most databases a thesaurus-based search is by default "exploded" to include all terms lower in the thesaurus hierarchy. This inevitably increases the number of retrieved records and often reduces the specificity of the search. Also, there is a time lag between the date an article is published and the date it is indexed. Consequently, when only thesaurus terms are used in a search, the most recent relevant articles might not be retrieved. Finally, previously indexed articles are often not retroactively modified to reflect changes made to the thesaurus. Thus, articles published before a certain term was added to the thesaurus will not be retrieved when that thesaurus term is used in a search. In summary, though the use of thesaurus terms might result in the retrieval

Clinical Question:			
In women over 40 with dilated cardiomyopathy, does treatment with warfarin lead to reduced morbidity and mortality due to thromboembolism?			
PICO-format:			
Patient	**I**ntervention	**C**omparison	**O**utcome
Women > 40 years with dilated cardiomyopathy	Warfarin	Not Applicable	Thromboembolism
DDO-format:			
Domain	**D**eterminant	**O**utcome	
Women > 40 years with dilated cardiomyopathy	Warfarin	Thromboembolism	

Table 1 | Example of a clinical question formulated according to PICO and DDO format.

of relevant articles that would otherwise have been missed, it might also decrease the specificity of a search and should always be accompanied by a search for free text terms or terms occurring in title and abstract only. This can be achieved by adding so-called "field tags" to the selected search terms. An example of a field tag in PubMed is "[tiab]", which limits searches to the title and abstract fields. Adding the field tag [tiab] disables automatic mapping of the search term to the thesaurus. In addition, it prevents searching in fields like "author name" and "affiliation", thereby increasing specificity.

An additional way to increase the specificity and sensitivity of a search is to use a pre-existing search filter as part of the search strategy. Search filters enable a search to be focused on a specific type of question (such as diagnosis, prognosis or treatment) or type of research. For example, in EBM, searches are often limited to randomized controlled trials using a search filter. Search filters can be designed either by subjective selection of search terms (10) or by following a more objective approach, using word frequency analysis and statistical analysis (11). In both cases, the performance of a search filter is usually tested on gold-standard sets of known records. Most filters are developed for use within a specific database and thus contain database-specific components, such as thesaurus terms and field tags. PubMed Clinical Queries (12) offers a series of EBM filters, as well as a filter for systematic reviews.

Databases

It is important to consider the databases in which the formulated search strategy will be executed. Although PubMed, with over 19 million references from 5200 biomedical journals, is often used as the starting point for literature searching in the biomedical sciences (13), there are other databases that should be taken into account if a search is to be as complete as possible. EMBASE (14), for instance, covers more pharmaceutical and pharmacological research than PubMed, and indexes more European journal titles. However, whereas the use of PubMed has been free since 1997 (15), other databases are often only available commercially, so they will not be freely accessible to everyone. Moreover, database-specific terminology in a search strategy, like thesaurus terms or database-specific filters, often complicates the use of one search strategy across different platforms. This also limits the use of meta-search engines like TRIP (16) and SUMSearch (17). Meta-search engines allow multiple databases to be searched simultaneously, but such search tools are, almost by definition, less refined than those available in the individual databases themselves.

Meta-search engines allow multiple databases to be searched simultaneously

An alternative to searching the primary literature (i.e. original research papers) is to make use of aggregated evidence, like systematic reviews. The Cochrane Collaboration (18) compiles systematic reviews on a wide range of therapeutic clinical questions, based on a broad search for randomized controlled trials (RCTs) in PubMed, EMBASE and non-indexed journals and conference proceedings. Searching the Cochrane Library (19) for both systematic reviews and RCTs is free, and free full-text access to Cochrane reviews has been made freely available in some countries, including the United Kingdom. In a commercially interesting field like EBM, a range of commercially available sources of aggregated evidence has come to market. Some examples are BMJ Clinical Evidence (20) and UpToDate (21). Products like these aim to offer the best available evidence, based on a

comprehensive and current overview of the literature. The value of these sources of evidence depends to a large extent on the quality and transparency of their literature searches. It should also be clear to users whether the source's recommendations are based solely on systematic review of the literature, or whether expert opinion also plays a role.

Text Mining and Semantic Search

All search techniques mentioned so far are based on finding specific terms within a database, either directly, through the use of a thesaurus or filter, or by using the results of searches conducted by third-party sources. With the exponential growth of scientific information and the possibilities of web publishing, there is growing interest in different ways of searching using text mining techniques. In text mining, free text is searched for patterns, rather than individual search terms, to retrieve high-quality information. One approach, building on the proposed architecture of the semantic web (22), focuses on the relationship between terms, thus taking into account their contextual meaning. Searching for relationships, rather than individual search terms, could enhance search specificity. Recently, an experimental biomedical search engine was launched that makes use of semantic search algorithms (23). In addition, text mining could be used to improve cross-linking of information in biological databases – for instance, those containing gene and protein sequences – with evidence in the literature (24). There is ongoing debate in the field as to whether text mining should be facilitated a priori – for example, by letting authors prepare structured abstracts in a computer-readable format (25) – or whether is should be done a posteriori using natural language processing (26). Finally, a triangle of publications, databases and end-users can be envisaged, with users annotating information in a Web2.0-like fashion (24). An experimental database using the latter concept is WikiProteins (27), currently incorporated into WikiProfessional (28).

In a changing information landscape, the role of the academic library is also shifting

Open Access

Knowing that information exists (for example, finding a citation in PubMed) is of little significance if the information itself is not available (i.e. the full-text article cannot be accessed). Currently, only 15% of citations in PubMed contain a link to the free full-text article (29), either through the publisher directly, or through PubMed Central, the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life-sciences journal literature (30). Over the last few years, the Open Access publishing model (31,32) has been gaining momentum, resulting in a growing number of open access journals (32) and institutional repositories. Many publishers are developing business models to accommodate Open Access – some examples are BioMed Central (33) and Springer Open Choice (34). Moreover, various funding agencies – like the National Institutes of Health in the U.S., the Wellcome Trust in the United Kingdom, and the Canadian Institutes of Health Research – now require that publications describing agency-funded research be made available through PubMed Central (35-37). Therefore, access to free full-text scientific articles will likely continue to improve.

Role of Academic Libraries

In a changing information landscape, the role of the academic library is also shifting. In the academic medical library, the stacks with bound issues of scientific journals have all but

disappeared. Often, the resulting free space is used to create a study environment with individual work stations and facilities for group work and discussion. One of the main roles of the academic medical library today is to provide a digital portal to scientific information. On a library's website, multiple search engines can usually be accessed, including commercial search engines to which users would have no free access otherwise. Often, libraries provide a direct link from search engines to the full-text articles in journals to which the library subscribes. Apart from offering access to databases and full-text journal articles, the academic medical library also plays an important role in teaching information literacy skills, enabling their users (including students, researchers and medical practitioners) to search the scientific literature effectively. The Association of College & Research Libraries (ACRL), a division of the American Library Association (38), has developed standards for information literacy in science (39). Several libraries have developed information literacy programs that aim to follow ACRL-standards (40,41).

Conclusions

The ability to carry out an efficient literature search is a key skill for students and researchers faced with an abundance of online scientific information. Future developments in text mining techniques, as well as the Open Access movement are expected to further increase the amount of scientific information that will be available to end users. Academic libraries continue to play an essential role in the dissemination of scientific literature and the teaching of information literacy skills. Taken together, these developments will greatly benefit students and researchers in locating the scientific information they need. H

References

1 NCBI. PubMed. Available at: http://www.ncbi.nih.gov/pubmed. Accessed April 20, 2010.

2 NCBI. PubMed. Available at: http://www.ncbi.nih.gov/pubmed. Search string: "2010/04/01" [Create Date] : "2010/04/15"[Create Date]. Accessed April 20, 2010.

3 Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. *Br Med J* 1996;312(7023):71-2.

4 Sackett DL. Clinical epidemiology: What, who, and whither. *J Clin Epidemiol* 2002;55(12):1161-6.

5 Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club* 1995;123(3):A12-3.

6 Dunning J, Prendergast B, Mackway-Jones K. Towards evidence-based medicine in cardiothoracic surgery: Best BETS. *Interact Cardiovasc Thorac Surg* 2003;2(4):405-9.

7 Rovers MM, van der Heijden GJMG. Translating research evidence into action in daily practice. *Otolaryngol Head Neck Surg* 2010;142(1):29-30.

8 NCBI. MeSH Database. Available at: www.ncbi.nih.gov/mesh. Accessed April 20, 2010.

9 Grobbee DE, Hoes AW. Meta-Analyses. In: Clinical epidemiology: principles, methods, and applications for clinical research. Sudbury, Massachusetts: Jones and Bartlett Publishers; 2009.

10 Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ* 2005;330(7501):1179.

11 Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. *J Med Libr Assoc* 2006;94(2):130-6.

12 NCBI. PubMed Clinical Queries. Available at: http://www.ncbi.nlm.nih.gov/sites/pubmedutils/clinical. Accessed June 10, 2010.

13 Haines LL, Light J, O'Malley D, Delwiche FA. Information-seeking behavior of basic science researchers: Implications for library services. *J Med Libr Assoc* 2010;98(1):73-81.
14 Elsevier. Embase. Available at: http://www.embase.com. Accessed April 20, 2010.
15 NCBI. NCBI News - August 1997. Available at: http://www.ncbi.nlm.nih.gov/Web/Newsltr/aug97.html. Accessed April 20, 2010.
16 TRIP Database. Available at: http://www.tripdatabase.com. Accessed April 20, 2010.
17 UT Health Science Center San Antonio. SUMSearch. Available at: http://sumsearch.uthscsa.edu. Accessed April 20, 2010.
18 The Cochrane Collaboration. Available at: http://www.cochrane.org. Accessed April 20, 2010.
19 The Cochrane Library. Available at: http://www.thecochranelibrary.com. Accessed April 20, 2010.
20 BMJ. Clinical Evidence. Available at: http://clinicalevidence.bmj.com. Accessed April 20, 2010.
21 UpToDate. Available at: http://www.uptodateonline.com. Accessed April 20, 2010.
22 World Wide Web Consortium. W3C Standards: Semantic Web. Available at: http://www.w3.org/standards/semanticweb/. Accessed April 20, 2010.
23 Quertle. Available at: http://www.quertle.info. Accessed April 20, 2010.
24 Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, *et al*. Text mining for biology - The way forward: Opinions from leading scientists. *Genome Biol* 2008;9(SUPPL. 2):S7.
25 Gerstein M, Seringhaus M, Fields S. Structured digital abstract makes text mining easy. *Nature* 2007;447(7141):142.
26 Hahn U, Wermter J, Blasczyk R, Horn PA. Text mining: Powering the database revolution. *Nature* 2007;448(7150):130.
27 Mons B, Ashburner M, Chichester C, Van Mulligen E, Weeber M, den Dunnen J, et al. Calling on a million minds for community annotation in WikiProteins. *Genome Biol* 2008;9(5):R89.
28 WikiProfessional. Available at: http://www.wikiprofessional.org. Accessed April 20, 2010.
29 NCBI. PubMed. Available at: http://www.ncbi.nih.gov/pubmed. Search strings: all [sb], free full text[sb]. Accessed April 20, 2010.
30 NCBI. PubMedCentral. Available at: www.ncbi.nlm.nih.gov/pmc. Accessed April 20, 2010.
31 Subarsky P. Open Access Publishing: Prelude to a paradigm shift in the dessimination of scientific literature. *Hypothesis* 2004;2(1):8-10.
32 Directory of Open Access Journals. Available at: http://www.doaj.org. Accessed April 20, 2010.
33 BioMed Central. Available at: http://www.biomedcentral.com. Accessed April 20, 2010.
34 Springer. Springer Open Choice. Available at: http://www.springer.com/open+access/open+choice?SGWID=0-40359-0-0-0. Accessed April 20, 2010.
35 NIH. National Institutes of Health Public Access Policy. Available at: http://publicaccess.nih.gov/policy.htm. Accessed April 20, 2010.
36 Wellcome Trust. Open and unrestricted access to the outputs of published research. Available at: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Open-access/index.htm. Accessed April 20, 2010.
37 CIHR. PMC Canada: Making Canadian health research accessible to all. Available at: http://www.cihr.ca/e/40259.html. Accessed June 10, 2010.
38 American Library Association. Association of College and Research Libraries. Available at: http://www.ala.org/ala/mgrps/divs/acrl/about/index.cfm. Accessed June 10, 2010.
39 Association of College and Research Libraries. Information Literacy Standards for Science and Engineering/Technology. Available at: http://www.ala.org.proxy.library.uu.nl/ala/mgrps/divs/acrl/standards/infolitscitech.cfm. Accessed June 10, 2010.
40 Rempel HG, Davidson J. Providing Information Literacy Instruction to Graduate Students through Literature Review Workshops. Issues in Science and Technology Librarianship

2008;53. Available at: http://www.istl.org/08-winter/refereed2.html. Accessed June 10, 2010.

41 Winterman B. Building Better Biology Undergraduates through Information Literary Integration. Issues in Science and Technology Librarianship 2009;56. Available at: http://www.istl.org/09-summer/refereed1.html. Accessed June 10, 2010.