

*Péter Jacsó**The University of Hawaii at
Manoa
jacso@hawaii.edu*

SAVVY SEARCHERS DO ASK FOR DIRECTION

Savvy searchers know that one of the secrets of doing a focused search is to use the controlled subject vocabulary of the database. They also know that a single database is unlikely to provide the perfect result for an in-depth literature search. The logical next step is to run a search using descriptors across several databases. But here comes the problem. Only a few online systems allow cross-database searching (DIALOG, DataStar, STN and from 1999 again WebSPIRS 4.0). However, they don't provide much help in finding the preferred terms of the controlled subject vocabulary.

DIALOG offers the closest to nirvana when it comes to simultaneous multiple database searching, but it stops short exactly in the subject searching area. True, it is possible to expand the thesaurus in the databases that have their thesauri implemented on DIALOG. But the number of such databases is disappointingly low, and you need to hop through the databases and do the thesaurus expansion in each to learn the exact descriptors. In addition, while DIALOG databases have a large number of prefixed indexes (also known as additional indexes, such as LA=language, DT=document type) that collocate together index entries generated from the same field, there is no DE= or ID= index, so the descriptors from the controlled vocabularies of databases appear in the Basic Index and may not stand out from index entries generated from the title, abstract and full-text fields. (There are exceptions because data in the Related Term column next to an index entry and/or the appearance of a compound term indicate that the term is a descriptor or an identifier, i.e. a term that is on tenure track to becoming a descriptor after a few years).

DIALOG - which magnificently helps users to discover the monstrous variations of journal, company and product names through its Journal Name Finder, Company Name Finder and Product Name Finder databases, which aggregate the index entries from the primary database indexes - has never ventured into developing a Descriptor Finder database.

What's the problem with descriptors of multiple thesauri?

The first problem is that there is a large number of semantic and syntactic variations for the same concept in the various thesauri. Is it nonverbal communication, nonverbal communication, metacommunication, meta communication or body language? Is it archeology or archaeology? Is it depression after birth, puerperal depression, postnatal dysphoria, postnatal depression or postpartum psychosis? It may be any of those depending on the database(s) being searched. Sometimes more of them are or were used in the lifetime of a single database. Would a savvy searcher know all these? Hardly. A casual user such as a college student? No way.

Would a thesaurus provide guidance? Sometimes yes, for a single database. But even the good quality Thesaurus of Psychological Indexing Terms would not lead you from wife abuse to its preferred term, partner abuse, because the former just does not exist even as a non-preferred term in the thesaurus. It does have, though, a cross-reference to partner abuse from spouse abuse, which is, by the way, spousal abuse in PAIS, and abused wives or abuse of wives in some other 'politico-linguistically' less correct databases. Would the overall intelligence of the patron help? Yes - for some of the patrons, some of the time. I keep forgetting, for example, that gender differences are sex differences in most of the controlled vocabularies, even if it is not as unambiguous a term. But my editor mildly reminds me that this may be a problem only for an ESL. To this I respond: like myself and the few million other users with English as a Second Language who use databases through online services.

Has anyone offered a solution yet?

KnowledgeFinder from Aries Systems was the first to offer an impressive solution for mapping the user's vernacular into the controlled vocabulary used in databases of the National Library of Medicine, and later into the Thesaurus of Psychological Terms of the American Psychological Association. Its very good natural language search capabilities are based on grammatical and statistical analysis of a large corpus of international medical literature. SilverPlatter offers

The screenshot shows a search results interface. At the top, it says 'Search Results'. Below that, it states 'Your Search "gender differences" found 9040 ITEMS'. There is a 'Refine' button and a link to 'use the Advanced Search screen'. A horizontal line with 'OR' in the center separates this from the next section. The section is titled 'Try these options for improving your search results:'. Underneath, there is a 'Search Words' section. It shows 'Occurred 0 times' and two radio buttons: 'Subject' (unselected) and 'Any field' (selected). To the right, there are two more radio buttons: 'Must include' (unselected) and 'May include' (selected). A list of suggested descriptors follows, each with a checkbox: 'gender differences' (checked), 'human sex differences' (unchecked), 'sex differences' (unchecked), 'sex stereotypes' (unchecked), and 'sexuality' (unchecked).

Figure 1: Suggested descriptors from various thesauri matching the user's term.

term mapping (known as the Suggest feature) for most of the databases it hosts, but the procedure is not tuned to the particular characteristics of the databases. The feature was withdrawn for two years but was reinstated in 1999. SilverPlatter can do cross-database searching but the Suggest feature is disabled during such searches. Ovid has a more sophisticated mapping algorithm that excels in finding quasi matches with thesaurus terms and does a very good job of extracting the most relevant descriptors assigned to articles that include in the title and abstract fields terms that appear in the user's query. But Ovid does not offer multiple database searching yet - let alone mapping into multiple thesauri. Enter KCL, the KnowledgeCite Library (www.knowledgecite.com), a new online service with nifty features.

What is so good about KCL?

There are too many features to list in this column, so let me just mention three of them. The KCL software does a natural language search and ranks the results by presumed relevance. The relevance rules are documented in the help file and represent a really good combination. They take into account such factors as term location, term proximity, document currency, term completeness, term frequency and term weight. The users have the chance to use traditional Boolean and proximity operators, as well as symbols made popular by Web search engines such as the '+' and '-' signs to indicate that a term must or must not occur in the item to qualify for retrieval. Truncation symbols may be applied to retrieve different variants of a root term. Exact phrases may be specified using a pair of quotation marks around the term. The searches may be limited to a particular data field either by using a prefix or using the pull-down menu in the advanced menu mode. It is similar to the solution offered by those few systems that allow cross-database searching. However, KCL goes one - or rather, several - step(s) forward from this point. If the user requests it by pressing the Search Suggestions button, it will list thesaurus terms that match the user words or the phrase, such as "gender differences", in some segments (entry term, cross-reference, related, broader or narrower term, definition and/or scope note) of all the thesauri built into KCL, or the ones in the discipline (say, Social Sciences) that the user chose (Figure 1).

By pressing the More button additional descriptors will be displayed. This process can be repeated as long as there are additional descriptors. The full

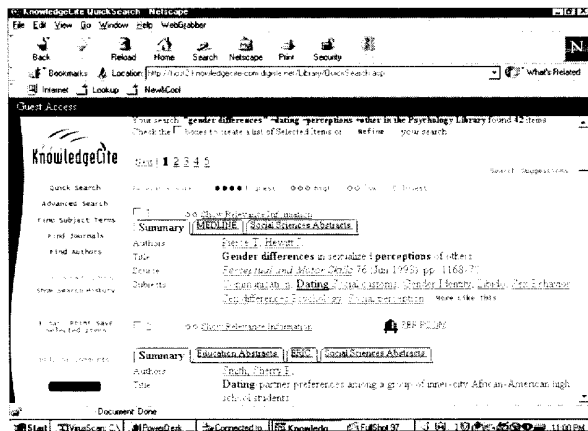


Figure 2: Multiple records for the same article from different databases.

entries for the descriptors can be displayed quickly. At the bottom of the thesaurus entries there are also cross-references to other thesauri. The rank ordering of the descriptors - as discussed in an upcoming conference paper - are not perfect in my opinion, but the deficiencies can be corrected easily. Alternatively, the users may invoke the Find Subject Terms option directly and navigate through the thesauri with the utmost ease without looking at the citation retrieved.

Looking at records retrieved, by the way, is also excellently implemented. If more than two databases have abstracts for the same document then both can be displayed one after the other by clicking on the tab (Figure 2). This grace and ease in the display of duplicate records I have seen only in those databases of NISC (National Information Services Corporation), which are built by integrating various databases in the same field of specialisation, as in the Wildlife Worldwide database. This is a combination of four databases featuring composite records. It avoids duplication among files by synthesising the content of two or more records, eliminating the redundant parts and retaining the unique information in each. In KCL the behind-the-scenes wizardry of term mapping is summarised in a chart that lists how the record scored by the various relevance criteria (Figure 3).

The third outstanding feature of KCL is how one can find journals covered by the different databases

that are relevant for a topic. DIALOG Journal Name Finder is excellent as long as you correctly contemplate the zillions of intra-database and inter-database abbreviations and acronyms, and if the journal name indeed includes the search term or its variation. The beauty of KCL is that every journal is assigned a Dewey Decimal code. So, if the user enters the term "women's studies" it will list highly relevant journals

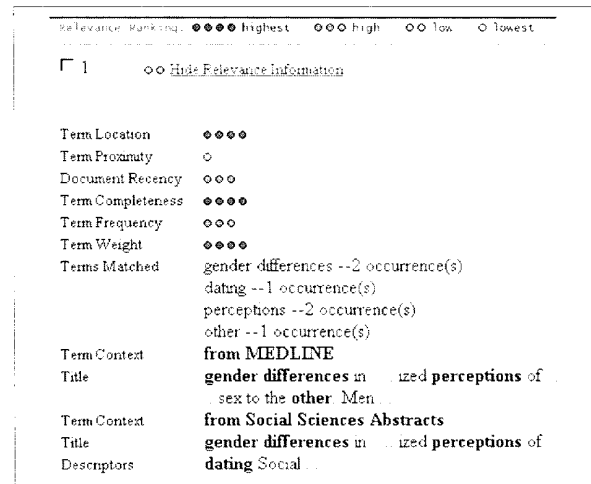


Figure 3: A record's relevance score for the query 'gender differences'.

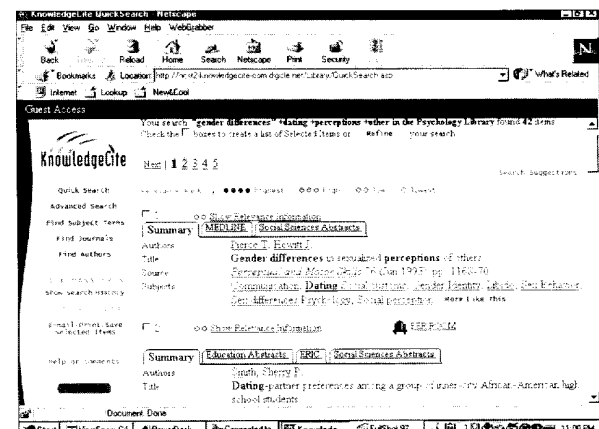


Figure 4: List of women's studies journals.

1. Jacsó, Péter. Cross database searching on the Web with term mapping from multiple thesauri. *Proceedings of the Twentieth National Online Meeting*, Medford, NJ: Information Today, Inc. (in preparation).

irrespective of whether the search term appears in their title or subtitle in one way or another (Figure 4).

In the beta version that I used there were only half a dozen thesauri and a good dozen databases but by

the time you read this the sources will certainly increase as content providers and users alike realise what an effective tool KCL offers.

© Péter Jacsó

The author

Péter Jacsó

Péter Jacsó is an associate professor of the Information and Computer Sciences Department at the University of Hawaii. He is the recipient of the 1998 Pratt-Severn/ALISE National Faculty Innovation Award for his innovative use of information technology in curriculum design and coursework. He is a frequent speaker at national and international conferences. His columns are published in *Database*, *Information Today*, *Computers in Libraries* and *Link-Up*. He has received various awards for his writings, including the 1998 Louis Shores/Oryx Press Award of the ALA Reference and User Services Association for his discerning database reviews.