DATABASE SEARCHING TIPS: PART 2

By Marvin Hunn

Part 1 reviews some basics elements of searching a structured, controlled-vocabulary metadata-based database using explicit operators and search syntax. It is practical and oriented toward procedures and techniques. Part 2 is a little more theoretical in that it focuses on understanding why search engines tend to have certain limitations. It warns you about problems related to discovery systems, to automatic processing by search engines, and to fundamental linguistic barriers to searching for concepts.

Discovery Systems

How are students to deal with the multitude of databases? It can be hard to remember which database covers which topic. It can be hard to remember differences between search engines. It can be tedious and time consuming to search three or four different databases. Wouldn't it be better to search many sources at once? That's the driving force that created the clamor for "single search."

Single search systems enable the researcher to search for resources from many different sources (databases) with a single search statement. There are two main ways to do this. One approach uses intermediate software to translate the search into statements compatible with the different search engines, send commands to all the different systems, and collect the results. This approach is called **federated** searching, broadcast searching, distributed searching, or cross database searching. This approach to single search began in the 1970s (I think). Around 2010 it was replaced by another approach which collects the data (records or articles) from all the different vendors and indexes them in one combined database so there is only one centralized index. This harvest-and-index approach produces a "unified **discovery**" system. It is indeed better than the federated search approach. This is what OCLC WorldCat aspires to be.

At its best, single search simplifies the complexities of searching. There is only one interface to master, and it appears there is only one database to search. People who use single search often use it as their sole means of searching library databases. And that is bad.

There are many problems with single search. Single search may insist on searching databases irrelevant to your current needs. This lowers precision. Single search normally does not support browsing, cross references, or specialized database features. It may display many duplicates (same article from several different databases.) In the case of the federated search approach, it limits you to simple search operators supported by all the databases, and it is slow because it must wait to retrieve results from many different intermediate searches before combining results. Unified discovery systems, on the other hand, require the vendor to normalize data from disparate sources into a common schema used by the discovery system. This homogenization usually means loss of distinctive or specialized fields (like the ATLA scripture field, for example). This is a significant limitation.

Perhaps the biggest problem with both federated and discovery approaches to single search is that **it masks the need to customize search terms to match the vocabulary of each specific database**.

Remember we are still talking about controlled vocabulary databases, not natural language databases. To understand the need to match the vocabulary of each specific database, let's compare subject terminology related to depression (the mood disorder) in three different databases that cover counseling.[1]

Aug-2015 searches of **subject** fields (not keyword)				
	search statement (default settings)	WorldCat (subset)	EBSCO Medline	EBSCO PsycINFO
1	depression	376,369	124,355	129,518
2	"depression emotion"	1,014	1	**22,414**
3a	"mental depression"	699	0	27
3b	"depression mental"	**17,597**	5	27
4a	"depressive disorder"	**88,155**	**79,872**	6,643
4b	"major depression"	4,629	424	**95,623**
5a	Prozac	185	9	64
5b	fluoxetine	**10,181**	**7,909**	**3,720**
6a	"serotonin uptake inhibitors"	**17,054**	**16,411**	31
6b	"serotonin reuptake inhibitors"	1,949	146	**4,622**
6c	ssri (abbrev for Selective SRI)	2,812	243	416
6d	ssri OR "serotonin uptake inhibitors" OR "serotonin reuptake inhibitors" OR "serotonin re-uptake inhibitors"	20,474	16,712	4,723

First examine lines 1-4. In WorldCat, "*depression*" refers to many different things including economic slowdown and the mood disorder. So it lacks precision. 3b shows "*Depression, mental*" is a more precise heading for the mood disorder. Medline uses "*depression*" for mild/temporary mental depression, and "*depressive disorder*" for severe/chronic conditions. PsycINFO uses "*depression (emotion)*" and "*major depression*" to make the same distinction between minor and major depression. Prozac is a drug used to treat depression. The generic drug fluoxetine is a better search term in all three databases when we limit to subject fields. This drug is a member of the class of serotonin reuptake inhibitors. In lines 5-6 we see the importance of using "*serotonin uptake inhibitors*" in Medline verses "*serotonin REuptake inhibitors*" in PsycINFO.

Different databases use different terminology. Sometimes they use the same terms in different, even contradictory, ways. A search statement that works well in one database may work poorly in another database. Automatic translation between subject headings in different databases could in theory mitigate the problem. Often there is not a predictable one-to-one equivalence between terms in different databases, so automatic translation is difficult, and that may be why it is rarely offered at present.

[1] Medline covers technical medical journal literature, including psychiatry. PsycINFO covers scholarly psychology broadly, including clinical psychology. WorldCat covers popular and scholarly works in all subject areas.

Variation in terminology between databases is a very serious problem. It is a fundamental linguistic problem (as are synonymy and polysemy in general).

Another important problem with discovery systems is that they tend to contain a lot of sparse records. Why? Discovery systems are under pressure to be comprehensive. But they can't get all the metadata they need. Some A&I creators (abstracting and indexing companies that create the metadata, such as PsycINFO or ATLA) will not lease their metadata to discovery vendors. These A&I companies want searchers to use the native A&I search engine and interface so searchers will recognize where the metadata is coming from. Name brand recognition helps justify a high price for the product. So discovery systems are turning to other sources to obtain metadata. For example, since they cannot get metadata from PsycINFO, they get sparse metadata from journal publishers for many of the same journals that are indexed in PsycINFO. Sometimes A&I owners will license abridged records. They create rich, full metadata records (including subject headings and abstracts) for use in their own system, but lease sparse/abridged versions of that metadata (lacking subject headings or abstracts) to discovery vendors. Sparseness of records is a common and significant shortcoming in discovery systems.

So discovery systems usually have a mix of full records with controlled vocabulary subjects, and sparse records with natural language titles but no subject headings. Searching a discovery system with such mixed records can be difficult. Our introduction to using WorldCat discusses this issue at length (http://library.dts.edu/wc-intro).

In spite of these limitations, single search has its place. Understand its limitations, benefit from it, and know when not to trust it. Think about how you search. Single search products do often produce "good enough" searches in a fast, convenient manner. But don't let single search lure you into lazy research. Don't let this shortcut short-circuit your thinking and your education. Exclusive reliance on the convenience of single search is a symptom of lazy research. For the undisciplined, single search supports the myth of easy research, reinforces lazy habits, and facilitates uncritical use of sources. Resist these temptations. Use single search but don't misuse; use but don't be beguiled into sloppy work.

Automatic Processing

Sometimes a search engine will automatically do things we didn't ask (or expect) it to do. For example, the software might automatically perform pluralization (looking for both singular and plural forms) or stemming (looking for words that match the root or stem of a search term) or synonym expansion or substitution (automatically looking for additional synonyms). These automatic features are sometimes un documented. And they may produce unexpected results.

On balance, you will probably benefit from various automatic features. But such features should never operate in secret, and they should always be optional.

- A search engine should always tell you what it has done. If the system does not tell you what it has done, you cannot easily understand why you got the results you got, and it can be difficult or even impossible to revise and improve the search. Unfortunately, some vendors consider such

automatic features to be proprietary secrets that give them some advantage over competitors. They provide little or no documentation of the features so competitors cannot copy them.

- A search engine should always allow you to take control so you can help it or bypass it as needed. Few vendors offer a way to turn off automatic features. They have too much confidence in their own software.

Concepts, Not Words

Ordinarily we want to search for concepts, for certain semantic meanings. But search engines search for strings of characters (words, partial words, phrases, etc.) They don't understand what the strings/words mean. The difference between a concept/meaning and a word/string is very important. Consider the following.

- One string of characters can represent multiple different meanings (concepts). The mismatch between concept and string is evident in homography (e.g., a word spelled 'bank' referring to a financial institution verses a different word spelled 'bank' referring to the sloping edge of a river) and polysemy (e.g., the word 'late' meaning 'after the appointed time' and the same word 'late' meaning 'dead' as in 'the late Mr. Smith').[2]

- Two different strings can have the same meaning (e.g. 'anger' and 'wrath'). This is synonymy.

- One concept/meaning can be expressed in many different ways, using many different words and phrases. The concept/meaning can even be expressed in a novel or original way. Figurative language makes this all the more complex.

So there is often not a one-to-one correspondence between a string and a concept. Search engines use three main approaches to deal with this problem.

1. The classical approach is to search for strings. This is how the EBSCO and WorldCat search engines work. The classical approach may also consider other textual clues (like paragraphs, capitalization, headings and bold face font, etc.) It may also use controlled vocabulary and cross references from one term to another (broader term, narrower term, related term). In modern systems, the cross references may be enhanced with a "linked data" database which specifies relationships between words like 'this is a member of that' or 'this is a synonym of that.' The relations support inferences to construct a "knowledge graph."

[2] Technically, homographs are two unrelated words that happen to be spelled the same way. They may have been incorporated into English from two different languages, for example. A polyseme, on the other hand, is one word with multiple different senses or meanings. Often, there is a single original meaning to which the other meanings logically relate and from which they historically derive. You will see a lot of homographs in your Hebrew lexicon, and the distinction between homograph and polyseme will be important if you have to trace a Hebrew root through cognate languages. But the distinction between homograph and polyseme does not matter for our purposes. They are both 'one string, many meanings.'

2. The linguistic approach first analyzes a document based on a model of how language works. For example, to determine what might modify what, the software identifies parts of speech, phrases, etc. It has a dictionary arranged by "concepts" (not really). It attempts to discern the intended concepts and searches for them. The linguistic approach uses all the elements of the classical approach, too. Our models of how language works are primitive and this approach does not (yet) work well. It is unlikely you will encounter a search engine of this kind.

3. The statistical ranking approach uses the classical approach to generate a list of matching items, then applies calculated weights to rank order the results. Recently the statistical approach has been paired with artificial intelligence learning methods. One way to train the search engine is to feed it ratings (by people) for thousands (even millions) of searches. The ratings indicate precision and recall. The software learns to improve ranking of documents in this fashion. In some variations, the software may also learn to expand the search by identifying other ways of expressing the search statement. So the statistical approach starts with the classical approach, improves ranking, and may also evolve from there. It is the "evolutionary" aspect—the ability to learn and improve—that is called artificial intelligence. But the software does not "understand" language; it only uses trial and error to find new ways (rules) to get closer to matching the training data people provide.

The main point here is that search engines have not overcome basic linguistic problems. Artificial intelligence techniques are producing ever better results, but do not naively trust the results. Do not search passively. Instead, actively review results and think about how to improve results until they are satisfactory. The search engine lacks your knowledge of language of the topic you are searching for. Use your knowledge to improve your searches.