DATABASE SEARCHING TIPS: PART 2

By Marvin Hunn

Part 1 discussed how to search a structured, controlled-vocabulary metadata-based database using explicit operators and search syntax. Here in Part 2 we continue that discussion.

Single Search, Synonymy, and Polysemy

We begin with a problem. 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 behind 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 was plagued with many problems: slow response time, pesky duplicate records from different sources, etc. It has been replaced by another approach which collects the data (records or articles) from all the different vendors and indexes them in one combined database so there is only one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 (Discovery with a capital D). It is indeed better than the federated search approach.

At its best, single search is quick and convenient. It 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. You can experiment with an incomplete search single implementation by picking "all databases" in EBSCO. (This would be incomplete because it includes only EBSCO databases, not WorldCat, ProQuest, JSTOR, Sage, etc.)

In spite of this popularity, single search very often produces unsatisfactory results. The main problem is that **single search 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]

[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.

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 term fluoxetine is a better search term in all three databases when we limit to subject fields. 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. Variation in terminology between databases is a very serious problem for single search. It is a fundamental linguistic problem (as are synonymy and polysemy in general).

There are other problems with single search. Single search may insist on searching sources/databases you don't want. Single search normally does not support browsing, cross references, term explosion 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. In the case of unified discovery systems, some vendors simply refuse to allow harvesting of their data, so single search will not be comprehensive. Also, unified

systems 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 a scripture field, for example).

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. It is all too easy to search "everything" (every database, every website) with a single naïve search statement, find some highlighted snippets of text in the first 10 items retrieved, copy those snippets (possibly even without even reading the context!!!!!), and finish the "research" paper as quickly as possible without any serious thought. 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 poorly documented. And they produce unexpected results.

Here is an example of a database-dependent automatic feature EBSCO documented in response to a question from a puzzled searcher.

> **Why do truncation (*) searches sometimes return fewer results?**
>
> Truncation searches using an asterisk (*) allow the EBSCO search engine to expand the query into multiple possible keywords. In some circumstances using the * operator can return fewer results. For example: The search "pediatric* and nursing" vs "pediatrics and nursing". The truncated query returns 4322 results while the non truncated query returns 5525 results.
>
> **The root of the issue comes from a behind the scenes feature of the EBSCO*host* search engine that applies a generic default thesaurus to searches (i.e. different forms and tenses of a word, even British spelling alternates).**
>
> As a rule we have never applied default thesaurus terms to truncation searches. This is why some phrases can return fewer hits. In this specific circumstance when a user searches for (pediatrics) we also search for the alternative spelling (paediatrics) found in the default thesaurus. However when the user searches for (pediatrics*), we never consult the thesaurus so "paediatrics" is not retuned and you get fewer results. The reason we have not used the default thesaurus for

truncation searches is that the stem could have too many possible keyword combinations with which to apply the thesaurus after they are generated.

In the coming months we will explore enhancing this functionality: If the stem of the truncation search is in the default thesaurus, then apply synonyms. Heart attack* would search "myocardial infarction" but heart attac* would not. pediatric* would search "paediatrics" but pediatri* would not.

Please note, this behavior does not apply to all EBSCO*host* databases. For example, because the CINAHL default thesaurus does not employ plural and possessive forms, this behavior does not apply to truncated searches in those databases.

ID: 4151

Topic: Interface Features, Database Products and Content

Link: http://support.ebsco.com/knowledge_base/detail.php?id=4151

Updated: September 2014

Automatic features are designed to make searching so easy you don't have to learn how to search. And they work for some searches and needs. 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. Unfortunately, few vendors offer a way to turn off automatic features. They have too much confidence in their own software.