DATABASE SEARCHING TIPS: PART ONE
By Marvin Hunn

The “Searching WorldCat” and “Searching ATLA” exercises were designed to make you think about how to search a structured, controlled-vocabulary metadata-based database using explicit operators and search syntax.¹ The main practical points are summarized here in Search Tips Part One. At the end of Part One we will briefly touch on a key theoretical linguistic problem with search performance. In Search Tips Part Two we will turn our attention to a newer approach to search and retrieval, based on keyword searching of documents using natural language input and document ranking. And we will pay more attention to theory.

Four Generalities

1. Picking good search terms is more important than using complex search statements. This is the key to success.

2. Many databases use standardized terminology like Library of Congress Subject Headings to express subject content. They are sometimes called “descriptors.” This is the "controlled vocabulary" (in contrast to “free-text” or "natural language") approach to expressing access points. One concept can usually be expressed in many different ways. Standardized terminology brings order to this chaos by selecting one term (a word or phrase) for each concept. This is usually achieved by creating a thesaurus that lists and defines all officially endorsed terms, with definitions and cross references to/from related terms and unused synonyms. When you browse a list of subject headings or names in ERIC, for example, you are looking at a controlled vocabulary list. Web pages and documents normally use natural language, not standardized terminology, so Internet search engines like Google use natural language.

3. Initial searches can often be improved. Evaluate the results of every search for possible improvements and then search again. Refine. Persist. We may refer to this as using feedback to improve searches, or as recursive searching or as successive searches.

4. Precision is inversely related to recall (completeness). A very precise search will not be comprehensive or complete; a very complete search will not be precise—it will include many, many irrelevant citations.² This inverse relation is especially true for systems that search natural language full-text documents (like Internet search engines and full-text journal article databases), but it also holds for systems that search standardized terminology. Because of this inverse relation, you must sometimes choose between either high precision or high recall, and formulate your search statement accordingly. Don’t frustrate yourself trying to formulate a perfect search which is 100% precise and 100% complete.

¹ Confused by terminology like “metadata” and “controlled vocabulary?” If so, see http://library.dts.edu/Pages/RM/Helps/glossary.shtml as background.

² Precision and recall (completeness ) are usually defined as follows. If $d =$ number of documents retrieved, and $R =$ number of relevant documents in the database, and $r =$ number of relevant documents retrieved by your search, then precision% = $r/d$ and recall% = $r/R$. 
How to Find Good/Better Search Terms

Selecting good search terms is extremely important. Sometimes you will need to find additional synonyms in order to formulate a more comprehensive search. At other times you will need to change terminology to formulate a more precise search. Listed below are some tactics to help you with both goals when you encounter problems. If you are searching a controlled vocabulary database, then focus on finding the correct standardized terminology for that particular database. If, on the other hand, you are searching a natural language database that lacks vocabulary control, then synonym discovery tactics and use of phrases will be more important.

1. **Execute successive searches and use feedback recursively to improve searches.** Conduct preliminary searches in the database. At this preliminary stage you may search for “obvious” subject terms or you may search for relevant specific titles you already know about. **Review results** of every search for relevant new terminology. Note words used in titles. Note terminology used in subject headings. **Execute new searches** with the new terminology. For example, an initial search for “pastor” retrieves some records. Examine those records and notice many of them use the LC term “clergy.” Include “clergy” as a search term. The successive searches approach can be used in any searchable database, even those that lack subject headings. This approach is just common sense, but it is commonly neglected because searchers are not willing to patiently and thoughtfully review results.

2. **Browse headings.** Many databases will let you browse term lists. Unclear on the distinction between searching and browsing? Searching allows you to look for any arbitrary word or combination of words anywhere in a record. Browsing is fundamentally different. It allows you to scan a sorted list of fields in various categories (author, title, subject, etc.). It is a two step process. Ordinarily you supply a word or words you expect to match the start of a field (e.g., the first two or three words of a title). The database then displays a sorted list with the closest match at the start of the list. Then you pick one or more specific entries from the list, and matching records are displayed. So browsing lets you see a list of potentially useful terms.

3. **Use cross references.** But many databases lack explicit cross references.

4. **Consult LCSH/thesauri.** Check the Library of Congress subject heading guide or some other relevant list of subject headings. ERIC, PsycINFO, MEDLINE and some other databases have their own thesauri similar to LCSH that include cross-references. Consulting LC is helpful even if the particular database you are searching uses something else as its list of standardized terminology. The “subject” icon in the FirstSearch version of WorldCat will let you query LCSH. VIAF.org is the most comprehensive multi-lingual, multi-disciplinary authority file.

5. **Brain storm.** Still don't have the magic terms? Step back and define what you need. List the distinct concepts. List synonyms for each concept. Remember to include central Bible passages. Record all of this in a master list of terms. At this point also take a minute to ask yourself what your goal is. What do you want: high precision or high recall? (See below

---

This successive searches approach is sometimes called “pearling” (where the record you start with is the pearl, and you grow a string of pearls with successive searches based on subject headings in those records) or “following the tracings” (where subject headings and other access points in a catalog record are known as tracings).
discussion on how to retrieve more or less, which parallels how to increase recall or increase precision.)

6. Read before you search any more. Read an overview of your topic in an encyclopedia or textbook. Note additional terminology. Note bibliography at end of article. READ any highly relevant works you have already identified and THINK about what is really relevant. Continue to follow footnotes. (This processing of using footnotes/references in one known work to find other useful works is sometimes called citation chaining.) Update your master list of search terms based on anything you read. Try your searches again.

7. Try a different database or specialized bibliography.

8. Evaluate terms one by one. If you have many terms, and you are trying to do a very thorough search in large database, then it may be worthwhile to execute individual mini-searches on each term or concept. You will discover some terms are unhelpful and should be discarded. Make the terms fit the specific database. The terminology you are using may fail in one database, but succeed in another.


How to Retrieve More

When your search retrieves insufficient material, take the following steps.

1. Check your spelling/typing first. Spelling errors are the main cause of “zero hit” searches (searches that retrieve nothing).

2. Specify additional synonyms. Think broadly. They may be subject headings, title phrases, or call number ranges (in WorldCat, for example). You can even use specialty publishers as proxies for subject terms. What you do will depend on the specific situation. Sometimes there is an obvious solution; often there is not. The section on “How to find good/better search terms” tells you how to find synonyms. “OR” synonyms together.

3. Be less restrictive. As needed, relax logical, proximity, and field constraints in your search statement.

4. If your database searches brief bibliographic records (author, title, subject) by default, but has an option to search the full-text of articles, try that option.

5. Consider using a different database or a print bibliography. The library has over 1500 specialized paper bibliographies.

How to Retrieve Less and/or Increase Precision

Some databases (like WorldCat) are so enormous that they will yield thousands or even millions of records for a general search. You can almost always limit the search to recent publications in English. Would just a few recent works meet your needs? More often you need to increase precision, not just limit the results. And that can be a difficult task.

1. If your initial set is the result of ANDing terms, consider using proximity operators or searching for exact phrases to increase precision. For example, “church and history” as a subject yields about 20,000 in the library catalog, but the exact phrase “church adj history” retrieves half as much. Of course both of these searches are very broad.
2. In many databases, a subject search may not be limited to subject headings. For example, default “subject” searches in many databases include words from both subject headings and titles. If your initial search was a keyword search that included both subject words and title words, and there is a subject heading that closely matches what you need, then re-execute the search using the exact subject heading, and limiting it to subject headings only. Nearly all databases provide a way to do this. Your revised search will retrieve fewer items and may increase precision. However, in my experience there is often NO standardized subject heading that accurately reflects the concept. Sometimes there is a stock phrase used in titles or abstracts or in the full text of the articles that does accurately reflect the concept. In such a case a title search for that phrase would be better than a search of the subject heading. Subject headings are not always the way to increase precision.

3. If there are still too many citations resulting from a subject search, then further reduce the set by requiring some of the same key terms appear as both subject and title words. This will probably also increase precision. If you are searching a free-text database that indexes every word in the documents, then limiting to title words almost always increases precision while decreasing recall.

4. If this set is still too large, and you can think of additional precise terms, then require that some of these terms also appear in the final set. In other words, AND another term or set to your previous set. Repeatedly add additional terms until it is a manageable size. This is sometimes called the “successive fractions” approach.

5. Further limit the topic in a new way. For example, to the initial search of “liberation theology” you might choose to AND “Latin America” to further limit the topic.

6. Abandon that database for the moment. Find some other source that will provide a short list of “best books” on the topic. How? Consult the bibliography in a textbook or the bibliography in an encyclopedia article or other reference book.

**“Single Search” and a Bit of Theory**

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 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. The other approach 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 or “vertical search” system. It is better than the federated search approach.

Single search is quick and convenient. It masks the complexities of searching. There is only one interface to master, and it appears there is only one [logical] database to search. People who use
single search rarely complain about it. Indeed, they often love it. 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 low precision, low recall searches. The main problem is that single search masks the need to customize search terms to match the vocabulary of each specific database. To understand the need to match the vocabulary of each specific database, let’s compare terminology in three different controlled vocabulary databases that cover counseling.

<table>
<thead>
<tr>
<th>search statement (kw index)</th>
<th>Medline</th>
<th>WorldCat</th>
<th>PsycINFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 depression</td>
<td>229,776</td>
<td>157,252</td>
<td>159,427</td>
</tr>
<tr>
<td>2 “great depression”</td>
<td>104</td>
<td>17,301</td>
<td>172</td>
</tr>
<tr>
<td>depression n2 emotion</td>
<td>118</td>
<td>134</td>
<td>20,366</td>
</tr>
<tr>
<td>3 mental n1 depression</td>
<td>614</td>
<td>19,397</td>
<td>667</td>
</tr>
<tr>
<td>4 &quot;depressive disorder&quot;</td>
<td>69,023</td>
<td>6,078</td>
<td>11,650</td>
</tr>
<tr>
<td>5 “major depression”</td>
<td>14,750</td>
<td>3,041</td>
<td>75,413</td>
</tr>
<tr>
<td>6 Prozac</td>
<td>320</td>
<td>1,577</td>
<td>239</td>
</tr>
<tr>
<td>7 fluoxetine</td>
<td>9,464</td>
<td>1,453</td>
<td>4,953</td>
</tr>
<tr>
<td>8 serotonin uptake inhibitors</td>
<td>14,032</td>
<td>496</td>
<td>542</td>
</tr>
<tr>
<td>9 serotonin reuptake inhibitors</td>
<td>5,794</td>
<td>1,034</td>
<td>7,080</td>
</tr>
</tbody>
</table>

2011 stats

Compare lines 1-5 of the table. In WorldCat, “depression” often refers to economic slowdown. “Depression, mental” is the LC heading for the topic we are interested in. Medline uses “depressive disorder” for severe conditions, but “depression” for milder conditions. And the keyword “depression” appears often in a non-psychiatric, non-counseling context like “depression of the immune system” so it is important to limit searches just to the subject field. PsycINFO uses “depression (emotion)” and “major depression” to make the same distinction between major and minor. In lines 6-9 we see the generic drug name fluoxetine is a better search term than the brand name Prozac in Medline and PsycINFO. We also 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.

There are other problems. 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

---

4 Duplicate records may be the most common complaint. Single search typically returns multiple records for the same item from different databases.
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, we expect single search will become pervasive. 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.