11-15-1987

Searching By Controlled Vocabulary or Free Text?

Carol Tenopir
University of Tennessee - Knoxville

BY CAROL TENOPIR

Searching by Controlled Vocabulary or Free Text?

ONE OF THE most common mistakes made by new online searchers who are *not* librarians is to rely completely on free-text, natural-language search strategies. One of the most common mistakes made by new online searchers who *are* librarians is to go to the opposite extreme and use only controlled vocabulary descriptor searching. Even experienced searchers often wonder when it is best to use free-text techniques and when it may be best to use descriptors. Unfortunately, there is no clear-cut or easy answer. After examining the results of over 20 years of research studies and looking at search strategies from searchers of all levels of experience, the only answer I can give to the free-text vs. controlled vocabulary dilemma is *it depends*. It depends on: 1) the database; 2) the vocabulary itself and the indexing policies that dictate how it is applied; 3) the topic to be searched; and 4) the requester.

The database

Many databases don't even have controlled vocabulary indexing, but the bibliographic databases used most frequently in libraries almost always do. According to research conducted by Martha E. Williams (and summarized in my column, "The Database Industry Today: Some Vendors' Perspectives, *LJ*, February 1, 1984, p. 156-157), the most frequently used databases in libraries are (in alphabetical order):

- ABI/INFORM
- BIOSIS
- CA Search (Chemical Abstracts)
- COMPENDEX (Computerized Engineering Index)
- ERIC
- Magazine Index
- Medline
- NTIS
- PTS (Predicasts) files
- PsycInfo (Psychological Abstracts)

All of these have a descriptor field that includes terms selected from some sort of controlled vocabulary or vocabularies.

Many of these databases use an extensive and carefully created thesaurus with a reputation for quality beyond the single database. For example, COMPENDEX uses "SHE" (Subject Headings for Engineering), ERIC uses "Thesaurus of ERIC Descriptors," Medline uses the excellent "MESH" (Medical Subject Headings), and PsycInfo uses "Thesaurus of Psychological Index Terms." ABI/INFORM developed a thesaurus and added controlled descriptors at the suggestion of users.

The other databases have controlled vocabularies that are typically used with less confidence or ease. Magazine Index has created a "Subject Guide to IAC" that uses modified and extended Library of Congress Subject Headings. NTIS records are indexed by one of several different thesauri depending on the subject and agency from which the document originated. CA Search uses a drug name authority and "Headings Lists" and Index Guides in place of standard thesaurus. BIOSIS relies heavily on codes and classification numbers that aid searching, but require training to use effectively. Predicasts has a company name thesaurus and a subject term listing.

More than one way

Although all of these databases have a descriptor field, the other fields available for free-text searching vary. Title words may always be searched, but only some databases have abstracts. Other fields such as captions or notes may serve as a short abstract-like text, but without the intellectual decision making that goes into the writing of an abstract.

The usefulness of titles for searching varies according to the subject matter of the database. Technical and research articles typically have longer and more meaningful titles than the popular articles found in a database such as Magazine Index. Short or ambiguous titles might be enhanced by the indexers (as Magazine Index does) but the titles are still usually shorter and less informative.

The meaningfulness of individual words in titles, abstracts, or other free-text subject-related fields varies with the subject matter. The level of ambiguity in language and the number of synonyms per concept have been shown to be subject dependent. Usually words in a so-called "hard" or "concrete" discipline (e.g., engineering or physics) are less ambiguous. Words in softer ("abstract") disciplines (e.g., philosophy or sociology) may be less meaningful. In soft disciplines the searcher must do a more careful job of complete synonym development in the search strategy. Studies have shown that a combination of title words and abstract words provide the most comprehensive (highest recall) free-text searches in all types of bibliographic databases.

The trade-off with the increased recall offered by abstracts is a corresponding increase in the number of false drops. If long abstracts are present and the database is large (over 500,000 records or so), abstract words can retrieve an unmanageable number of documents with an unacceptable number of false drops. Controlled vocabulary searching should be considered instead. Unless a search needs to be comprehensive, the best strategy in most large databases in this situation is probably a compromise of searching both title words and descriptors (and identifiers if they are available). Identifiers are used in some databases to add new subject terms or to control subject-related proper nouns.

The vocabulary

Not all controlled vocabularies are created equal. Before relying on a database's descriptors, the searcher should have a good feel for the quality and limitations of the individual controlled vocabulary and the

Carol Tenopir is Assistant Professor at the Graduate School of Library Studies, University of Hawaii, Manoa

policies that dictate how it is applied.

Most of the databases that use controlled vocabulary descriptors use a thesaurus from which indexers select appropriate terms. Thesauri offer searchers many advantages: control of synonyms, control of variant word-forms (e.g., singulars or plurals), control of homographs, and hierarchical term relationships that facilitate search strategy development.

Other databases may use only a term authority list. Such a list merely documents accepted word forms with references from unused forms. Unlike a thesaurus, there is no hierarchical arrangement of Broader Terms, Narrower Terms, and Related Terms to help the searcher (and indexer) with word and concept selection.

Ideally, a thesaurus takes much of the burden of synonym development and concept building from the searcher. Unfortunately this promise may not always be fully realized.

The ideal thesaurus

A thesaurus must be sufficiently specific to allow topics to be defined narrowly. Assigning the term "BIOLOGICAL METHODS" to an article on gene splicing (as one database did) offers little chance of precision. Free-text searching of titles and abstracts works better in this case.

On the other hand, if a broad concept needs to be searched, a good controlled vocabulary will facilitate this. Searching for articles on folk tales of any Southeast Asian country, for example, is much easier in a database where the controlled vocabulary takes care of the country concept by assigning a broad overall category of Southeast Asian nations in addition to the specific country names. At least the printed thesaurus should list all of the nations within a broader region so a searcher need not consult an atlas to prepare a search. MESH goes one step further by assigning classification codes to terms and allowing these codes to be searched at any level. For example, a search on the truncated code number for steroids will retrieve all of the steroid categories and specific steroids.

A thesaurus must be as up-to-date as possible or the database should have an identifier field where new topics may be added. "Latchkey Children" was an identifier in ERIC before it recently became a descriptor. Like "Downloading" and "Burnout," latchkey children is a very specific concept with an exact meaning.

A thesaurus should have enough cross references and broad, narrow, or related terms to allow a searcher to find all descriptors that pertain. Scope notes should appear in at least the printed version to clear up ambiguous terms. A "connectedness" ratio has been suggested as a way to evaluate a thesaurus. Connectedness is the ratio of terms linked with at least one other term to the total number of terms in the thesaurus. A connectedness figure of two to five has been recommended by researchers.

The ability to view the thesaurus online is helpful for search strategy development. Unfortunately, the capability to invoke a thesaurus automatically while searching online is rare. (Notable exception: WILSONLINE.)

The impact of indexing

In addition to thesaurus quality, the database's policy on indexing affects the success of controlled vocabulary searches. Who is doing the indexing will have a big impact as well. H.W. Wilson's experienced professional indexing staff has a reputation for quality. Other databases may have volunteer indexers or indexers spread out over such a wide geographic area that it is difficult to be consistent. Research has shown that in most databases there is little inter-indexer consistency. Finding all possible descriptors for a topic or using free text in conjunction with descriptors may thus be necessary.

Policies on the number of terms assigned will directly affect search results. If a database limits the number (e.g., "assign no more than five descriptors per document"), then only major topics of an article (or the topics seen by the indexer to be most important) will be covered. Relevant documents may be lost. On the other hand, assigning too many terms may negate the precision value offered by controlled vocabulary. The separation into major and minor descriptors (as ERIC does) can be a great help.

The topic

If one facet of a topic is very broad, using a descriptor term for this facet may be the best strategy. For example, in a search on gymnastic and tumbling programs in primary grades, the controlled term "Primary Education" will not only limit the number of citations retrieved, it will shorten response time considerably. Gymnastics and tumbling are easily free text searched and give better results when they are. A search in PsycInfo that includes the facet of personality factors, or a search in COMPENDEX with one facet of testing or design, will do the same.

If all facets in a search are of equal importance and likely to retrieve sets of equivalent size, controlled vocabulary may not be as important. Graffiti on the subway system, for example, is easily searched free text and retrieves a greater number of relevant items than descriptor terms in Magazine Index.

Controlled vocabulary is frequently useless for new ideas or jargon. If your topic is a current one (e.g. "designer drugs"), free text may be the only effective way to get relevant documents.

The requester

If the requester wants a "few good ones" (a high-precision, low-recall search) and there are descriptors that match their topic, then descriptor searching is usually the best bet. Controlled vocabulary searching has been shown to be the most cost-effective search method, allowing retrieval of a small set of precise items at a low cost per relevant record retrieved. Premenstrual Syndrome and (Migraine/de or headache/de) in the current Medline file on DIALOG will retrieve approximately a dozen highly relevant documents at a minimal cost.

If a more comprehensive search is needed, however, research has shown that controlled vocabulary searching alone is usually inadequate. Indexing policies, limitations of the vocabulary, and the nature of the English language all make it difficult to be comprehensive with descriptors. (Unfortunately, research has shown it is difficult to achieve complete recall even with free-text strategies.)

Summary

No one search method is the best in every database, for every topic, or for every requester. The ability to use a combination of controlled vocabulary and free-text techniques or change strategies online is the mark of a good searcher. Like so much else in online searching, the ability to judge which strategy is best comes with experience and intuition. Although research provides some possible answers, search strategy is not a science. Ultimately, it all depends on the searcher.