Who Needs Controlled Vocabulary?

By Raya Fidel

■ **Whenever they search a database, how do searchers decide whether to use a descriptor, a textword, or both? Observation of a total of 281 real-life searches performed by 47 professional searchers shows that although some searchers preferred descriptors and others textwords, the decision about which type to use depended on each specific situation. Searchers' reasons for search-term selection revealed a set of rules that guided their selection. The nature of a term and of a request, as well as the searcher's personal preference, contributed to the selection of search terms, but the number of databases required for a request, the availability and quality of the thesaurus, and the quality of indexing were the major factors to affect search-term selection.**

Many bibliographic databases provide access with two types of search terms: descriptors from a controlled vocabulary, and textwords for free-text searching. While some databases cannot be searched with descriptors because they do not have thesauri or indexing, almost all databases can be searched with textwords. In addition, most searchers use both descriptors and textwords during their professional careers.

Searchers' opinions about these types of search terms vary: some like to search with textwords, others prefer to use descriptors, and still others claim "it depends." These opinions are certainly based on professional experience in searching, but what is really known about the actual behavior of searchers and the thinking processes underlying their behavior? Do searchers generally use textwords more frequently than descriptors, or vice versa? Is the decision whether to enter a textword or a descriptor idiosyncratic, or is it based on rules? What are the reasons for the selection of search terms? Why do some searchers prefer textwords and others descriptors?

A study of online searching practices answered these and other questions about the selection of search terms. The study team (the author with the assistance of four graduate students) observed 47 professional searchers as they performed their regular, job-related searches of bibliographic databases. The searchers, who worked in different types of libraries and specialized in a variety of subject areas, were asked to think out loud, and their verbalizations of thought processes during the search were taped. Each searcher was observed for approximately five searches, for an overall total of 281 searches. In analyzing these searches, the study team identified each case in which a searcher selected a descriptor or textword and investigated the reasons that led to the selection. These analyses revealed general rules that explain the selection of search terms. Highlights from the study's findings are presented here, and a detailed description is available elsewhere.[1]

What Type of Search Term Is Best?

Even though some of the study's searchers preferred to use textwords and others descriptors, the answer is: *the decision about which type of search term to use depended in each case on the specific situation.* Various findings of the study substantiate this answer.

The Selection Routine

The most clear evidence for this answer is a decision tree, called the Selection Routine, that emerged from analyzing and consolidating all the search terms selected by the study's searchers. This Routine shows *when* the searchers decided to use textwords and *when* to use descriptors. In other words, it illustrates specific situations and the preferred choice in each of these situations.

For example, the Selection Routine describes what searchers did when the only descriptor they could find for a term was broader in meaning than the term itself. Consider a case in which the only descriptor that could be entered for **ducks** was **WATER BIRDS** (This example was created by the author for illustrative purposes and is not an actual search observed in the study. All such examples are marked with an asterisk).

Here, the choice was between entering the term as a textword or entering the broader descriptor. The Selection Routine explains that if the term was well defined and suitable for textword searching (as **ducks** most likely is), and if the user wanted just a few items about the topic, entering the term as a textword was probably the best choice. If, however, the term was not suitable for free-text searching, or if the user did not want to miss any item (high recall search), searchers entered the descriptor. In fact, the Routine shows that there was another option. If the user was interested in a high precision search, that is, the user wanted to see only a few highly relevant citations and recall was of no importance, some searchers used the combination of textword and descriptor, **ducks** and **WATER BIRDS**. This formulation retrieved only articles indexed under **WATER BIRDS**, that is, those that were *about* the birds, and selected among them those that mentioned ducks.

Another example from the Selection Routine is the case where searchers could not find a descriptor to represent a term. The most direct solution was to search the term as a textword, but there were other options: searchers could probe indexing or enter the term as a descriptor anyway. According to the study's searchers, it was best to enter the term as a textword when: (a) the patron wanted very specific retrieval, or (b) the term itself was very well-defined, that is, an "ideal" textword, or (c) the term appeared in titles and abstracts of *relevant* citations, or (d) previous attempts to use related descriptors resulted in poor retrieval.

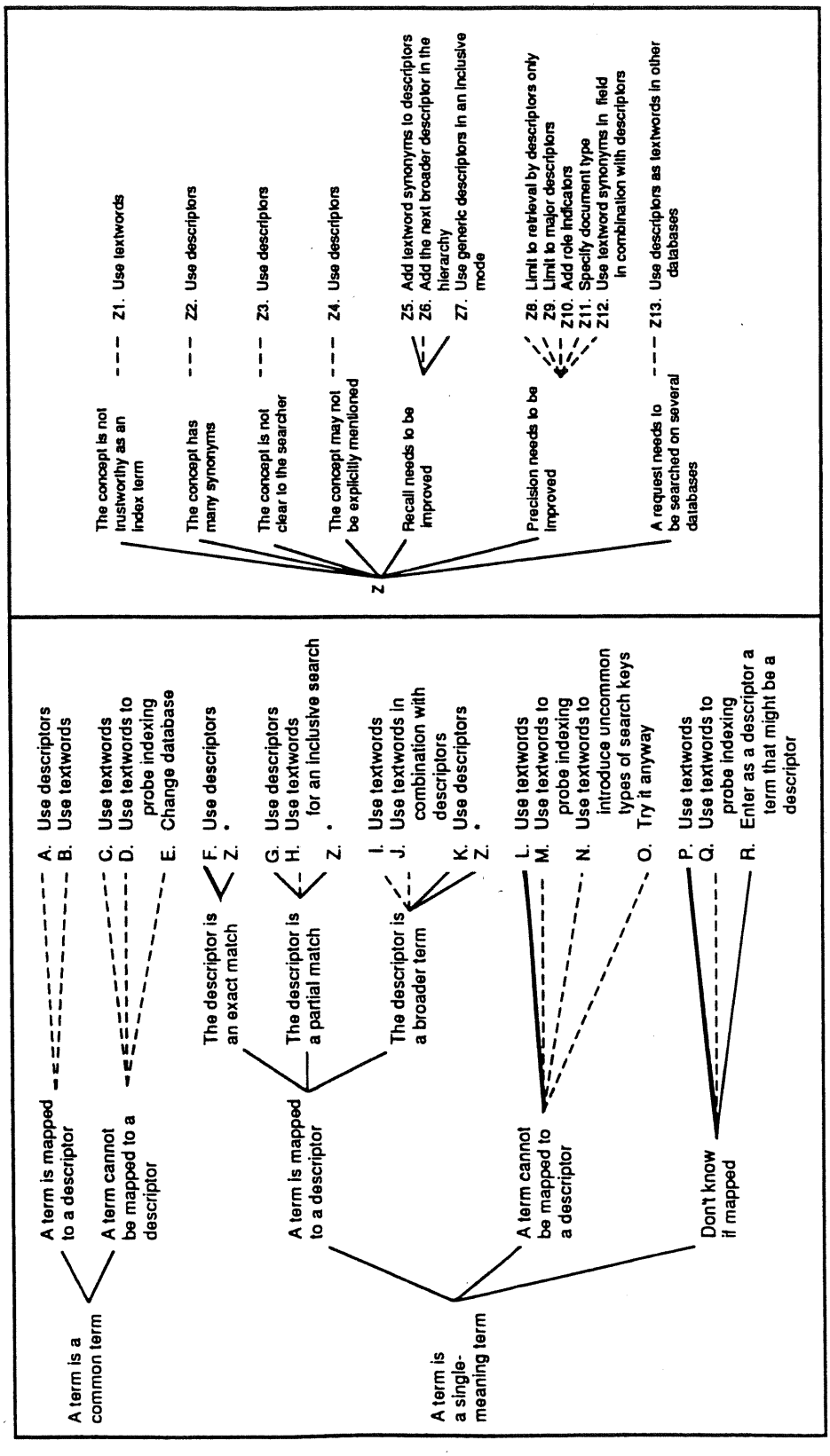
If none of these conditions existed, searchers probed indexing. They entered a combination of textwords and then examined the indexing of relevant citations retrieved. Consider a search about students' behavior during final examinations; the formulation of textwords **behavior** *and* **students** *and* **final** *and* **examinations** retrieved a few citations.* All the relevant citations had the descriptor **EDUCATIONAL TESTS**, thus indicating that it was a useful descriptor. In addition, if a term was new and had no descriptor, searchers found it beneficial to enter it as a descriptor, just in case it had recently been added to the thesaurus. Similarly, if a term was a descriptor in a number of databases, searchers sometimes entered it as a descriptor without knowing if it was a descriptor in the current database. Among the study searches, entering terms as descriptors that way was successful more frequently than one would have expected.

The Selection Routine describes 21 situations of this sort, and thus provides a comprehensive array of reasons for the selection of search terms. It shows that term selection depended on whether or not a term was well-defined, on how closely the descriptor described the term, on the quality of the thesaurus and indexing, and on a host of other factors. It is beyond the scope of this article to describe the Routine. Figure 1 displays it and a detailed description is provided in the final report of the study.[1]

Which Type Is Used Most Commonly?

Another way to investigate which type of search term was preferred by searchers is to determine which one was used most frequently. If one type of search term was preferred, one would expect this type to be used more frequently than the other, given the large sample of searches. The study showed that of a total of

Figure 1


A term is a common term
A term is mapped to a descriptor
A term cannot be mapped to a descriptor
A. Use descriptors
B. Use textwords
C. Use textwords
D. Use textwords to probe indexing
E. Change database
A term is a single-meaning term
A term is mapped to a descriptor
The descriptor is an exact match
F. Use descriptors
Z. •
The descriptor is a partial match
G. Use descriptors
H. Use textwords for an inclusive search
Z. •
The descriptor is a broader term
I. Use textwords
J. Use textwords in combination with descriptors
K. Use descriptors
Z. •
A term cannot be mapped to a descriptor
L. Use textwords
M. Use textwords to probe indexing
N. Use textwords to introduce uncommon types of search keys
O. Try it anyway
Don't know if mapped
P. Use textwords
Q. Use textwords to probe indexing
R. Enter as a descriptor a term that might be a descriptor
Z
The concept is not trustworthy as an index term
Z1. Use textwords
The concept has many synonyms
Z2. Use descriptors
The concept is not clear to the searcher
Z3. Use descriptors
The concept may not be explicitly mentioned
Z4. Use descriptors
Recall needs to be improved
Z5. Add textword synonyms to descriptors
Z6. Add the next broader descriptor in the hierarchy
Z7. Use generic descriptors in an inclusive mode
Precision needs to be improved
Z8. Limit to retrieval by descriptors only
Z9. Limit to major descriptors
Z10. Add role indicators
Z11. Specify document type
Z12. Use textword synonyms in field in combination with descriptors
A request needs to be searched on several databases
Z13. Use descriptors as textwords in other databases

3,189 search terms, searchers selected 1,607 descriptors and 1,582 textwords. Thus, even though some searchers used more textwords than descriptors, and others selected descriptors more frequently, overall, it averaged out. To find out to what extent searchers' individual preferences affected the selection of search terms, we analyzed the reasons searchers gave when they selected search terms.

Reasons for the Selection of Search Terms

Their reasons can be divided into three groups: request-related, database-related, and searcher-related. **Request-related** reasons are those that stem from the nature of the specific request being searched. For example: "the request has a relatively large number of concepts," "high recall (or precision) is needed," "the term appeared in titles and abstracts of relevant articles," "the term was added while online," or "the term is specific and well-defined."

Examples of **database-related** reasons are: "I need to search several databases," "I don't trust the descriptors and/or the indexing," or "the thesaurus is not available."

Searcher-related reasons, on the other hand, are individual tendencies and beliefs that searchers used to explain their selection of search terms. For example: "I prefer to use descriptors," "the use of textwords increases recall," "if a term is well-defined, there is no need to use descriptors," "terms suggested by users are the best for retrieval," or "I prefer to start with textwords and then check for descriptors."

An examination of these categories revealed that the most dominant were database-related reasons (48% of all reasons), with request-related next (32%), and searcher-related reasons last (20%). That is, 80 percent of the time, searchers decided whether to enter a textword or a descriptor based on the attributes of the database they were searching or on the requirements of the request. Their personal preference for a particular type of search term played a very minor role.

These findings show clearly that the selection of search terms is based primarily on the specific situation and that no one type of search term is always superior to the other.

Why Some Prefer Textwords and Others Descriptors

Since no one type of search term is superior, why do searchers have personal preferences in this area? This question was not easy to answer. Several approaches proved unfruitful before the answer began to emerge.

Searchers' Reasons

To answer this question we looked first at the searchers' explanations for their selection of search terms. They offered three main arguments to explain a personal preference for one type of search term: (a) the use of textwords increases recall; (b) the use of descriptors increases precision; and (c) terms entered during online interaction are bound to be textwords. A close examination revealed, however, that actual searching behavior did not conform with these beliefs.

Searchers used either textwords or descriptors to increase recall, depending on the specific conditions. Similarly, precision was increased with textwords or with descriptors. For example, to increase recall some searchers added textword synonyms, but others entered a broader descriptor with all its narrower terms ("explode" or "cascade"). Similarly, at times they increased precision by using the descriptor, **EDUCATIONAL TESTS**, rather than the textword, **examination**; the textword combined with the descriptor, **ducks** *and* **WATER BIRDS**, resulted in higher precision than when the descriptor alone was entered.

To further test the prevalent belief that the use of textwords increases recall, we measured the correlation between two variables: "recall tendency" and "textwords ratio." Recall tendency was defined as the degree to which recall was of concern in a search. This was measured by the percent of search-strategy modifications that were aimed at increasing recall. Textwords ratio is the number of textwords divided by the total number of search terms.

The working assumption was that if searchers acted on the belief that the use of textwords always increased recall, then searches in which recall needed to be improved would show a relatively high use of textwords. The statistical test between these two variables, however, showed no significant correlation. These conclusive results emphasize the point that *both* textwords and descriptors are important for the improvement of retrieval, whether for higher recall or for better precision.

Another of the searchers' arguments for selecting textwords was that terms entered during the session at the terminal are bound to be mostly textwords. Here again, this belief is *not* supported by the results of the study. Searchers explained the selection of *both* textwords and descriptors with the reason that the term was entered while online.

Here too, a statistical test reaffirmed the conclusion. The working assumption here was that searches in which the search strategy was changed a relatively large number of times should include a higher textwords ratio than searches with only a few strategy modifications. The variables textwords ratio and number of strategy modifications, however, do not correlate. That is, while online, searchers may add textwords they spot in titles or abstracts, or they may use descriptors assigned to relevant citations.

These analyses show, then, that searchers' stated reasons for their preferences of search terms are not borne out by their actual searching behavior. Is it possible, then, that a preference for textwords or descriptors is a matter of individual searching styles?

Individual Searching Styles

To test this idea we defined several variables that describe individual searching behavior, such as the average number of search terms selected by a searcher per search, the average number of search modifications per search, the average number of search terms selected without consulting a thesaurus, the average number of databases searched per request, or the subject area in which a searcher specializes. We then tested the association between these variables and the average textwords ratio per search for a searcher. Such statistical tests can show what characteristics are common to searchers who prefer textwords and those common to descriptor searchers.

Results showed that most variables do not correlate with textwords ratio. For example, searchers who typically used more search terms than their peers, or those who modified their strategies most frequently, did not use textwords more frequently than others. There were, however, two exceptions: the subject area and the number of databases.

Subject Area

Statistical analyses showed that the variable of searcher's subject area correlates significantly with textwords ratio. On the average, the percentage of textwords used by searchers in each subject area is as follows:

Subject area	Textwords Ratio for a Searcher
Medicine	34%
Social sciences and humanities	39%
No subject specialty	57%
Science and technology	76%

Science searchers used textwords more frequently than any other group of searchers. At first glance, this finding seems obvious. It is commonly believed that searches in the scientific literature do not require the use of descriptors because the scientific terminology itself is specific and well-defined. But this argument is not a valid explanation for the finding in this case, given the difference between science and medical searchers. Medical terminology is scientific terminology, and there is no evidence to assume that it is less specific or well-defined than other scientific terminologies. Yet, science searchers used textwords more than twice as frequently as medical searchers.

This distinction between medical and science searchers becomes even sharper when one examines the rate of thesaurus neglect in those subjects, that is, the ratio of textwords entered *without* consulting a thesaurus. Like textwords ratio, the rate of thesaurus neglect correlates significantly with the subject area, and the averages are revealing:

Subject area	Rate of Thesaurus Neglect
Medicine	0%
Social sciences and humanities	13%
No subject specialty	29%
Science and technology	32%

No medical searcher entered a textword without checking a thesaurus first, but 32 percent of the textwords entered by science searchers were not checked in a thesaurus.

These results were interesting. They showed that science searchers used more textwords, but they did not support the most common explanation that the *reason* lies in the nature of the scientific language. Further explorations revealed that the reason for a preference of one type of search term lies partly in the number of databases that are usually required for a request and partly in the quality and availability of thesauri and indexing.

Number of Databases

Statistical analyses revealed that the average number of databases a searcher used per search correlates significantly with textwords ratio. That is, searchers who habitually searched several databases for a request used more textwords than searchers who used a single database. Similar association was found between the rate of thesaurus neglect and number of databases. This indicates that the larger the number of databases to be searched per request, the more likely it is for a searcher to neglect a thesaurus.

This result can partially explain the difference between science and medical searchers. On the average, a medical search required 1.33 databases, while a search of the scientific literature required 2.64 databases. That is, science searchers have to have used twice as many databases as did medical searchers, and therefore, they used textwords more frequently than their peers.

On the surface, the association between number of databases and textwords ratio seems almost trivial: a search that spans a number of databases is likely to include more textwords than descriptors because it is time consuming to consult thesauri and look for descriptors for each database. For the same reason, searchers who usually search a number of databases for each request are likely to develop a habit of neglecting to consult a thesaurus and of using more textwords than descriptors.

Because the number of databases seems to be an important factor in the selection of search terms, it is important to see that this factor is also one of the *reasons* for developing a personal preference for a certain type of search term. Searchers are free to choose whether to enter textwords or descriptors, but the number of databases to search for a request is determined by the distribution of information among the databases; it is a given. Undoubtedly, a searcher who prefers to use textwords would move from one database to another more easily than one who preferred to use descriptors. But even a textwords searcher would change databases only when it is required for the success of a search, while a searcher who habitually has to search a number of databases for each request would likely develop a habit of neglecting to consult a thesaurus and of preferring to use textwords.

The conclusion that having to search a number of databases may cause a searcher to neglect to consult a thesaurus and to use textwords is supported by additional evidence. Among the database-related reasons that were given for the selection of textwords, the reason "I had to perform a multi-database search" was mentioned 20 percent of the time.

Quality and Availability of Thesauri and Indexing

The findings thus far show that a person who habitually searches a number of databases is likely to neglect to consult a thesaurus and to use textwords. Does it mean, then, that a person who searches only a few databases is likely to consult a thesaurus and use descriptors most of the time? In other words, are there any other reasons why searchers prefer to use textwords?

To answer this question we examined the reasons given by searchers for neglecting to consult a thesaurus. There were 803 instances in which searchers neglected to consult a thesaurus. Of these instances, 57 percent (461 instances) were related to the database searched.

The database-related reasons given for neglecting a thesaurus are as follows:

Reason for Neglecting a Thesaurus	Times
Don't trust the thesaurus or the indexing	129 (28%)
The term would not be in the thesaurus	107 (23%)
Had to perform a multi-database search	117 (25%)
Had no access to the relevant thesaurus	108 (24%)
Total database-related reasons	461 (100%)

Thesaurus availability and quality are important factors in the selection of search terms. Thesauri of poor quality or limited availability encourage searchers to enter textwords without checking first whether or not appropriate descriptors exist. Thus, even a person who searches only a few databases may avoid consulting a thesaurus and may prefer to use textwords if the thesauri in question are of poor quality or unavailable.

To verify that the opinions searchers have about the "quality" of a thesaurus or indexing are solid, rather than just an individual searcher's idiosyncratic notion, we measured the textwords ratio and the rate of thesaurus neglect for each database. Because the databases in the study were selected by the study's searchers, rather than in a random fashion, we cannot generalize the results. Nevertheless, these ratios suggest the hypothesis that databases, their thesauri, and indexing acquire a "reputation" among searchers: Some are typically searched using descriptors, and for the others, those that are searched most commonly with textwords, searchers often do not bother to check the thesaurus. If proven valid, this hypothesis will show that some thesauri are highly regarded and heavily used, while others are unattractive and therefore commonly ignored.

Taken together, the results of this study show that preference for type of search term is affected partly by the searching experience of the individual. Medical searchers are the least likely to prefer textwords or to neglect to consult a thesaurus because most of the time they use one database which they perceive to be of high quality: MEDLINE. Thus, even when they search other databases, they first consult the thesaurus. On the other extreme are science searchers who, for each request, routinely search a number of databases and who work with some that do not have thesauri available or that have thesauri and indexing of poor quality. Social science and humanities searchers still use significantly more descriptors than do science searchers because of terminological difficulties and because of the limited number of databases they can search.

Who Needs Controlled Vocabulary?

Personal preference aside, the benefit of controlled vocabulary has been questioned numerous times.[2, 3, 4] Because the creation and use of controlled vocabularies is very labor-intensive, it has been claimed that thesauri are not cost effective. To date, however, no study has proven this claim.[5] While the study reported here does not prove the opposite (that is, that thesauri *are* cost effective), it clearly proves the importance of controlled vocabularies to searching.

The important role of thesauri and indexing is manifested by the rate at which searchers consulted thesauri and the rate at which they found and used descriptors. The study's searchers consulted thesauri for 75 percent of the search terms they selected; and 80 percent of the times they consulted a thesaurus, they selected a descriptor. Further, almost half of the times a descriptor was selected, it matched the request term exactly. In addition, as reported earlier, 50 percent of all the terms entered were descriptors. Thus, searchers made heavy use of thesauri and descriptors. Differently put, eliminating controlled vocabulary would have prevented the searchers from having a choice three quarters of the time and from entering their preferred choice of search term half of the time.

This popular vote for controlled vocabulary is reinforced by another of the study's findings. An important purpose of an index language is to control for synonyms. Using

textwords, searchers themselves are supposed to control for synonyms: each concept in a request should be represented by a number of synonyms ORed together. Study results indicate, however, that searchers who entered mostly textwords used, on the average, the same number of search terms as those who entered mostly descriptors. This finding suggests that often searchers who prefer textwords do not compensate for the lack of vocabulary control by using synonyms. This implies that while theoretically possible, synonym control during online searching is not practical. Controlled vocabularies, therefore, serve as the main vehicle for synonym control.

Because thesauri prove to be extremely beneficial to online searching, it is useful to examine what could be done to promote their use. The results of this study clearly show that better quality and availability as well as support for multi-database searching are likely to increase the use of controlled vocabularies. However, while some improvements could be introduced right away, others require additional research and development.

At present, there are no technical barriers to making thesauri more easily available. The availability of thesauri is totally in the hands of database producers and search-system vendors. Searchers will use thesauri if they can afford them. In reality, many printed thesauri are expensive, and searching them online is often more expensive than searching the database.

The quality of thesauri is a more complex issue. Some thesauri were put together sloppily and their deficiencies are easy to detect. For others, however, it is not known as yet what specific problems searchers have with them or with the databases' indexing. Because it is clear that higher quality standards in thesauri and indexing are required, it is important to begin studying specific deficiencies searchers find in existing systems. Only with such studies will it be possible to develop standards to improve the quality of thesauri and indexing.

Lastly, assistance in multi-database searching will be available when switching languages are developed and widely used. While still experimental, these languages facilitate the "translation" of the vocabulary of one thesaurus into another, and the vocabulary of a searcher into the vocabulary of a designated thesaurus. Having a switching language as a component of an "intelligent" interface, a searcher, using only a few keystrokes, would be able to display for each database the descriptors that represent a request term.[6] There are still unsolved problems in the construction of such languages, but a few large-scale projects are under way. The National Library of Medicine, for example, is constructing the Unified Medical Language System (UMLS).

To conclude, this study shows that both textwords and descriptors are necessary for quality searching. While several future developments could significantly improve retrieval from online databases, the most immediate improvement is for database producers and search-system vendors to provide easy, flexible, and inexpensive access to thesauri. ■

References

1 Fidel, R. *Extracting Knowledge for Intermediary Expert Systems: The Selection of Search Keys.* Final Report. Syracuse, NY: ERIC. (ED 314 059)

2 Dubois, C.P.R. "Free Text vs. Controlled Vocabulary; A Reassessment." *Online Review* 11 (no. 10): 243–253 (August 1987).

3 Fugmann, Robert. "The Complementarity of Natural Language and Indexing Languages." *International Classification* 9 (no. 3): 140–144.

4 Katzer, Jeffrey. "A Study of the Overlap Among Document Representations." *Information Technology: Research and Development* 1 (no. 4) 261–274 (October 1982).

[5] Svenonius, Elaine. "Unanswered Questions in the Design of Controlled Vocabularies." *Journal of the American Society for Information Science* 37 (no.5): 331–340 (September 1986).

[6] Chamis, Alice Y. "Selection of Online Databases Using Switching Vocabularies." *Journal of the American Society for Information Science* 39 (no. 3): 217–218 (May 1988).

Raya Fidel is an Associate Professor at the Graduate School of Library and Information Science, University of Washington. She teaches courses in information science, database design, indexing and abstracting, and construction of index languages. Her research focuses on online searching behavior.