DOCUMENT RESUME

ED 144 619 | IR 005 286

AUTHOR	Clement, Linda Lee
TITLE	Health Effects Profiles for Searching Selected Lockheed DIALOG Data Bases.
PUB DATE	Jun 76
NOTE	56p.; Master's Thesis, University of North Carolina at Chapel Hill
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	Data Analysis; *Data Bases; *Health Conditions; Literature Reviews; *On Line Systems; *Search Strategies
IDENTIFIERS	BIOSIS; Chemical Abstracts Condensates; DIALOG; Enviroline; National Technical Information Service; Pollution Abstracts

ABSTRACT

This preliminary study attempted to determine the most effective search strategies for the topic "health effects" in relation to specific chemicals and/or pollutants--in this case, asbestos--for each of five selected Lockheed DIALOG data bases: BIOSIS Previews, Chemical Abstracts Condensates (Chemcon), NTIS, Enviroline, and Pollution Abstracts. Strategies were developed, searches run, results calculated, and strategies synthesized for the larger data bases--BIOSIS, Chemcon, and NTIS. Strategies were developed for Enviroline and Pollution Abstracts by selecting possible search terms/codes from relevant citations, testing hypothetical search strategies for recall and precision, and then synthesizing these strategies. All citations from each data base search were evaluated for relevance and precision was calculated. Following data analysis, the study recommended basic health effects strategies for each data base. Trends ascertained in searching these data bases by testing with asbestos showed that codes work well in BIOSIS; Chemcon and NTIS require a combination of codes and words; words alone or codes and words together could be used in Enviroline, and Pollution Abstracts require all words whose identities can be pinpointed. A list of user aids, bibliography, tables, and illustrations are included. (Author/KP)

HEALTH EFFECTS PROFILES FOR SEARCHING SELECTED LOCKHEED DIALOG DATA BASES

by

Linda Lee Clement

A master's paper submitted to the faculty of the School of Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science.

Chapel Hill

June, 1977

Approved by:

Evelyn A. Morant
Adviser

LINDA LEE CLEMENT. Health Effects Profiles for Searching Selected Lockheed DIALOG Data Bases. A research paper for the M.S. in L.S. degree, June, 1977. 49 pages. Adviser: E. Moore.

This preliminary study attempts to determine the most effective strategies for health effects for each of five selected Lockheed DIALOG data bases (BIOSIS Previews, Chemical Abstracts Condensates, NTIS, Enviroline, and Pollution Abstracts) as the concept is used by the EPA library, Research Triangle Park, North Carolina, in relationship to substances. The effectiveness of different strategies for specific data bases are tested by determining recall and precision for an essential or core strategy plus that of additional strategies and comparing what the additional terms and/or codes did or did not add to the recall and precision of the essential strategy.

Tentative strategies were developed and summaries of areas of the searches still requiring testing are included. Definite trends can be established for each data base. Different strategies are required in each and levels of precision attainable vary with each data base. With the three larger data bases (BIOSIS, Chemcon, and NTIS), strategies were developed first, the searches run, results calculated, and strategies synthesized from the results. With the two smaller data bases (Enviroline and Pollution Abstracts), strategies were developed by selecting possible search terms/codes from relevant citations, testing hypothetical search strategies for recall and precision, and then synthesizing more final strategies.

Headings:

Online searching -- Strategies and profiles

Lockheed DIALOG data bases

Health effects -- Literature searching

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

The author would like to express sincere appreciation to Libby Smith, Head Librarian of the EPA Library, Research Triangle Park, N. C., and to Dr. Evelyn Moore for both their invaluable assistance in the undertaking of this project and for their suggestion that it be undertaken. Thanks go to Dav Robertson and Gail Gainey for their advice on style and content. Special thanks must be given to Dr. Herman H. Henkle who provided the author with the opportunity.

I. INTRODUCTION

One of the most important topics searched on computer data bases at the Environmental Protection Agency library at Research Triangle Park, North Carolina is that of health effects of substances. This topic is searched in conjunction with some specific chemical or pollutant or group of the same. Health effects is a broad, amorphous, and interdisciplinary category, varying in the aspects it covers and requiring numerous terms to be entered into the search strategy (profile). In addition, it is usually necessary to undertake the search in at least three, and even as many as eleven or more, data bases because of the interdisciplinary nature of the EPA's concerns. Therefore, it would be helpful if the effectiveness of different strategies for specific data bases could be determined in order to increase and assure accuracy, relevance, completeness (making certain important articles are not inadvertently missed), and to cut down on the cost of a search where additional, unnecessary terms can be excluded. The search results from such strategy testing should show terms that are essential plus additional terms that produce a pattern of increasing recall--possibly reaching a plateau--until an optimum level of recall and relevance is reached after which point the relevance will start dropping. So, it should be possible to determine which profiles produce the most desirable results.

Thus, the purpose of this study is to attempt to determine what are the most effective strategies for health effects for each of five selected Lockheed DIALOG data bases (BIOSIS Previews, Chemical Abstracts

Condensates, or Chemcon for short, NTIS, Enviroline, and Pollution Abstracts) used at the EPA library in Research Triangle Park, North Carolina. The results should be considered tentative since the strategies are regarded as preliminary until they can be tested with other pollutants and chemicals, especially those with health effects differing from those of asbestos.

II. REVIEW OF RELATED LITERATURE

Articles reporting a library's experiences with computerized literature searching are fairly common at present. Some deal primarily with operational costs (Calkins, 1977), others with the library users performing searches directly themselves or with training users (Shearer, 1975; Callaghan and Howden, 1972; Hines, 1975), while others give a general overview which covers a range of the searching experience (Prewitt, 1974; Schipma, 1974). Another common kind of article is the comparative type which frequently overlaps the experiential type (Laurence, 1974). These vary in what is compared. Some concern themselves more with comparing searching of different data bases (Beauchamp, 1973), some with the usefulness of different systems for retrieval (Verheijen-Voogd and Mathijsen, 1974; Prewitt, 1975), some do both (Weiss, 1976).

A number of online studies have been undertaken in EPA libraries. Calkins compared operational costs of different online systems with manual and batch searching (1977). Long and McCullough compared retrieval of different data bases as they related to environmental science search topics requested by researchers (1976; 1975).

Various combinations of topics appear in the numerous articles that have been written on searching computerized data bases. In addition to aspects already mentioned, some deal with in-house data bases, others with commercial ones; some with online searching, others with batch. The interest of this paper is a specific commercial online system, that is, Lockheed's DIALOG, and with health effects profiling for five of its data bases.

In the area of literature on search strategies and/or profiling, much less has been written for either online or batch systems, and little goes beyond the usual generalities of searching techniques that are included as only a portion of an article on computerized literature searching. (Literature searches on searching strategies/profiles on BIOSIS, Chemcon 3 and 4, ERIC, and NTIS plus a survey of searching articles indexed in the last five years of Library Literature and Library and Information Science Abstracts were used to gain this assessment.) Even less can be found when one considers what has been written specifically about DIALOG and other systems that incorporate full-text searching with more controlled techniques. Very few studies exist that devote themselves only to the indepth study of all the constituents involved in profile or strategy development. This may well be attributed to the cost involved in performing such studies. It cost approximately $310 in connect-time and printing of citations offline to perform the study presented in this paper. Several new publications devoted entirely to online information systems are beginning publication this year, e.g., Online, and may fill this gap. (While some of the data base manuals supply advice on profile development, and the manuals tend to be improving in quality and usefulness, they still do not supply searchers with much of the practical, individualized advice that is needed.)

While the "paucity of practical literature on the subject" of profile construction has been reported by Butterly (1975), it is fairly easy to find very general guidelines such as

> The major subtasks involved in framing a request include: 1) selecting search terms; 2) augmenting the terms with instances, synonymous phrases or other related terms; 3) expressing the logical relationships which exist between terms; 4) trying out aspects of the request of the whole request in order to discover how well it works; and, 5) explaining what to do with the retrieved records once the user is satisfied. It is generally agreed that a major advantage of interactive retrieval is that one can revise a request to conform to what one discovers about the data base. Some users are likely to carefully think out the request ahead of time and proceed one step to the next. Others are likely to skip right to the middle and add words to their request while they browse (Martin, 1975, p. 79).

or

> Many factors can influence the success or failure of a search. Primary causes of failures are lack of appropriate terms in our controlled vocabulary (some terms are too general and some too specific), lack of specificity in indexing or omission of necessary terms, and search formulations which do not adequately cover the request. Other failures are caused by inadequate user/formulator interaction (Jenkins, 1972, p. 425).

Others, in describing their searching cycle, also include free-text, iterative methods and the resultant revision of search strategies (Prewitt, 1974, p. 117) or interactive programs for preparing strategies as Schultz described for BIOSIS (1974, pp. 5-9). Sometimes review articles supply searching guidelines for a wide range of searching methods and/or systems (Stevens, 1974).

Some articles in an area closely-tied to profiling are those dealing with question negotiation in query formulation (Heim, 1975) and those dealing with indexing, its quality (Farradane and Yates-Mercer, 1973) and the effect of different methods on retrieval efficiency (Schipma, 1976).

Even those authors who provide more detailed guidance in search profiling are quick to point out difficulties in providing such directions. As Lynch says,

> Profile construction . . . is still largely a subjective and pragmatic process, depending to a considerable extent on skill and experience. For most users, it is intricate, time consuming and remote from their normal practices in consulting conventional sources (1974, p. 66).

Furthermore, according to Lancaster,

> It is usually difficult to make firm recommendations relating to search strategies. Nevertheless, we must carefully examine the failure analyses with a view to assembling a collection of pointers for searchers (1968, p. 157).

And, he does provide useful guidelines for both searching strategies and the concomitant failure analyses in several chapters in Information Retrieval Systems (1968, "Factors Affecting the Performance of an Information Retrieval System," pp. 64-78; "Analysis of the Test Data," pp. 130-150; "Interpretation and Application of the Test Data," pp. 151-159; and "Searching Strategies," pp. 198-207).

A summary of his guidelines is particularly germaine to this study and is as follows:

1) . . . a high level of exhaustivity of indexing makes for high recall and low precision. Conversely, a low level of exhaustivity of indexing makes for low recall and high precision (1968, p. 67).

2) . . . a highly specific index language will allow high precision capabilities in searching but will also tend to reduce recall performance. An index language of low specificity will tend to produce high recall figures but will not allow high precision performance (1968, p. 70).

3) Exhaustivity of indexing and specificity of index language govern the recall and precision capabilities of an index. However, the searcher is able to vary recall and precision performance for a particular search by the adoption of various searching strategies (1968, p. 70).

4) Given the ability to vary our search formulation (in order to retrieve more documents or fewer documents as the situation demands), by moving up or down hierarchies, by substituting synonyms, or by some other technique, we are able to carry out searches of varying degrees of generality. For any search, or group of searches, we can thus vary the

position at which we choose to operate on a hypothetical performance curve. Thus we can decide between sacrificing precision and going all out for a high recall performance, or sacrificing recall to obtain a high precision search; or we can adopt a compromise and operate somewhere in between (1968, p. 71).

5) Relevance standards of users of a retrospective searching system are obviously closely related to the generality of requests . . . [With] a very general request for the particular system being evoked [i]t should be possible to achieve both a high recall and a high precision figure for such a search since the requestor will accept any document that bears on the general subject . . . (1968, p. 78).

6) System failures attributable to indexing . . . are [of] two distinct types . . .: (a) those due to indexer errors and (b) those due to a policy decision regarding the average number of terms assigned in indexing. Indexer errors are themselves of two types: (a) omission of a term or terms necessary to describe an important topic discussed in an article, and (b) use of a term that appears inappropriate to the subject matter of the article. Omissions will normally lead to recall failures, while use of an inappropriate term (i.e., sheer misindexing) can cause either a precision failure (the searcher uses this term in a strategy and retrieves an irrelevant item) or a recall failure (the searcher uses the correct terms and a wanted document is missed because labeled with an incorrect term) (1968, p. 140).

In capsule form, the principal causes of information retrieval systems failure from indexing are: 1) lack of specifity, lack of exhaustivity, omission of important concepts, and use of inappropriate terms which lead to recall failures; and 2) exhaustive indexing and use of inappropriate terms which lead to precision failures. The principal causes of failure from searching are: 1) failure to cover all reasonable approaches to retrieval (e.g., not using one particular relevant term or term combination), too exhaustive formulation, and too specific formulation which lead to recall failures; and 2) not sufficiently exhaustive formulation, not sufficiently specific formulation, use of inappropriate terms or term combinations, and defects in search logic which lead to precision failures (Lancaster, 1968, pp. 150, 143).

More specific guidelines and examples to profiling exist, such as search term selection, different kinds of techniques in formulation, illustrative examples of why searches may fail (Lancaster, 1968, pp. 191-207); term relationships and word distance requirements and examples of search strategies (Lancaster, Rapport, and Penry, 1972, pp. 230-238); and profile construction in controlled-vocabulary data bases, free-text data bases, and interactive systems (Lynch, 1974, pp. 66-74). Scheffler describes a study in using Boolean NOT logic for improving SDI profile precision (1972), while Smith offers Venn diagramming as an aid in profile development (1976).

In the area of access points in searching, Williams emphasizes that

> Many research projects have analyzed the utility of various access points--terms in titles, abstracts, extracts digests, and controlled or uncontrolled index terms, key words and codes. The access points are evaluated with respect to recall, precision, and volume of material that must be checked by the user. One cannot generalize from such studies, because they are specific to certain data bases. The quality of indexing and abstracting varies among data bases, and the information content in titles varies among authors and among fields (1974, p. 233).

A few examples of access point studies are the Becker, Veal, and Wyatt comparison of efficiency when searching titles only, titles-plus-keywords, and titles-plus-abstracts in free-text chemical data bases (1972); the study by Lancaster, Rapport, and Penry on EARS (an epilepsy file) comparing searching on abstract (plus index terms) versus index terms alone (1972); the Fisher and Elcheson comparison of the effects of combining title words and index terms against using only either one of these accesses on the Nuclear Science Abstracts file (1972); and the Byrne evaluation of the relative effectiveness of searching titles, abstracts, and subject headings for a COMPENDEX data base (1975). Even though generalizations in this area can be risky, Byrne's assessment that "there is a general agreement that the addition of abstracts and/or other free-language words is beneficial with regard to recall" (1975, p. 224) more often than not is accurate, but one must realize that this addition can profoundly reduce precision. Roe, Micuda, and Seeds found that in searching the natural-language data base CAIN (now AGRICOLA)

> . . . success with title-word searching appears to vary inversely with the vocabulary size of a subject. The more limited, precise, or universal the terminology defining a subject, the greater the rate of success. Thus searches involving scientific names of unique processes were most successful in retrieving high percentages of relevant citations without the nuisance of "false drops" (1975, p. 796).

Lancaster and Fayen discuss searching using different methods of vocabulary control (1973, Chp. 11, "Vocabulary in the On-Line System," pp. 244-262).

Examples of articles providing information on searching techniques specific to the DIALOG system are two which also compare DIALOG with ORBIT. One of Weiss's concerns was searcher keystrokes. He also gave a thorough discussion of system commands (1976). Prewitt compared searching Chemical Abstracts Condensates on DIALOG and ORBIT and in the process covered searchable fields, subject searching, truncation features, searching techniques, and provided example searches (1975). Both these articles necessarily provide mostly general information.

Durkin and Smith discuss methods for retrieving environmental science-related information from BIOSIS Previews including use of the CROSS and Bio-Systematic Indexes and provide examples of search strategies (1975, pp. 15-16). They do not discuss the general concept of health effects, however. Nees and Green evaluated the BIOSIS data base, its indexes, and several systems for searching it including DIALOG (1976).

They concluded that "[i]nitial review of the Subject Guide to CROSS Index, CROSS Code, Biosystematic Code, and Guide to the Vocabulary of Biological Literature is essential" in strategy preparation (1976, p. 8). In addition to providing an example of a DIALOG strategy (1976, p. 33), they have a useful summary section including techniques and cautions for BIOSIS on DIALOG ("A. Searches for Regular Clients," pp. 16-25). Much of their observation and advice parallels that of searchers at the EPA library in Research Triangle Park, such as the powerfulness of the CROSS Code as a strategy tool (1976, pp. 21, 37) and the general necessity of limiting to major levels of CROSS Codes in most searches to prevent irrelevance and false drops (pp. 20-21).

III. METHODOLOGY

The description of this study and its results--thus the remaining chapters of this paper--necessarily assume experience with searching and are written accordingly.

The methodology in this study has two stages: 1) that involved in the development of the search strategies being tested and 2) that involved in the analysis of results from the strategies.

The methods used for developing the strategies are of two types. One was used for the three larger data bases--BIOSIS, Chemcon, and NTIS; the other for the two smaller data bases--Enviroline and Pollution Abstracts. For all data bases, however, the health effects will be those caused by asbestos. The stored asbestos strategy F87 that was ANDed with the other portions of each search appears in Figure 1. (The terms were selected from Standen, 1967.) This helps limit the number of citations

Fig. 1. Stored asbestos strategy F87.

S ASBEST?
S SERPENTINE
S CHRYSOTILE
S AMPHIBOLE
S ANTHOPHYLLITE
S AMOSITE
S FERROANTHOPHYLLITE
S CROCIDOLITE
S TREMOLITE
S ACTINOLITE
C 1-10/OR

to a manageable number and also determine relevance as asbestos has a wide range of health effect (Bogovski et al, 1973). Further control was obtained by limiting BIOSIS, Chemcon, and NTIS to a period of one year--those citations published in the 1976 volume(s) of the indexes; Enviroline to two years--1975-1976; and Pollution Abstracts to five years--1972-1976.

Methodology for Search Strategy Development--BIOSIS, Chemcon, and NTIS

Data and citations for these first three data bases were obtained by running one search on each. Each search was divided into several portions. In some cases the search had to be run in several steps because of storage overloading problems from the large sizes of some sets. The first portion contained terms and/or codes considered essential, ANDed with the limited (i.e., to one year) asbestos strategy. This appraisal was based both on strategies/profiles developed by the head librarian during several years of experience searching health effects and the author's own experience with searching. The rest of each search contains additional sections of terms and/or codes to increase recall. These were formulated from other strategies developed by the searchers at the library, thesauri and search guides for the data bases, and free-text words. Each section was ANDed with the limited asbestos strategy. Then the NOT function was used to determine what the additional terms/codes did or did not add to the recall and relevance compared with the essential strategy. See Figures 2, 5, and 6 for the strategies on BIOSIS, Chemcon, and NTIS. The profiles were not exhaustive because the expense would be prohibitive. However, manual methods were used to evaluate the results in more depth than could be afforded in online evaluation. Further discussion of the manual methods appears in "Methodology for Analysis of Results" and in "Data Analysis".

The important health effects that must be covered in a health effects strategy are toxicological, carcinogenic, mortal, and other pathological effects such as mutagenesis and teratogenesis. Both free-text and indexer-supplied terms and codes were necessarily used to achieve this end.

The primary focus of health effects is on humans, mammals, and mammalian experimental animals, therefore, fish and plants were excluded where such an exclusion was incorporated into an index code. And, since health effects are the focus, profiles did not limit their focus to specific organs or systems (e.g., cardiovascular diseases). It should be noted, however, that by using these general strategies, these "specifics" are picked up.

The expected optimum precision was that 70% of the citations would be relevant (not necessarily useful). This figure is the approximate percentage of relevance per data base found by Long in her masters paper (1976, p. 38) and is the level desired by the head librarian.

Methodology for Search Strategy Development--Enviroline and Pollution Abstracts

The approach used for these two smaller data bases was to obtain online all citations related to asbestos for the years given earlier--attempting to achieve a number of citations as close to 100 as possible--and work backwards in developing a health effects strategy or strategies. This method was chosen because of 1) the very small numbers of citations that would be found on health effects of asbestos which would make evaluation and calculation of relevance and precision shaky and 2) the specificity of the indexing would make it necessary to enter an exhaustive number of terms for even an "essential core" strategy, at considerable expense while producing little information on general terms to use.

By working backwards, the citations relevant to health effects could be assigned manually as with the method used for BIOSIS, Chemcon, and NTIS described later in "Methodology for the Analysis of Results"; then possible free-text and controlled terms could be identified among these and counted for frequencies; and finally, tentative strategies developed and tested for their recall and precision.

Methodology for Analysis of Results

The key measurement devices in the analysis of this study are precision, recall, and relevance. Precision is the ratio or percentage of relevant answers retrieved compared with the total number of references retrieved (Saracevic, 1975, p. 327; Verheijen-Voogd and Mathijsen, 1974, p. 141; Lancaster, 1968, p. 56). "In addition to the number of relevant references retrieved, their precision is considered to be a criterion for the effectiveness of a data base" (Verheijen-Voogd and Mathijsen, 1974, p. 141), or for the purposes of this study, a search term or strategy or portion thereof. Although Saracevic defines recall as "the ratio of relevant answers retrieved over the total number of relevant answers in the file" (1975, p. 327; see also Lancaster, 1968, p. 55), it will be used in this study to designate the number of citations retrieved whether relevant or not. It would be impossible, given the resources of money and time available for this study, to determine the actual number of relevant answers in files as large as the ones being studied. By comparing the different numbers of citations recalled with different profiles/terms and coupling this comparison with the years of experience in searching health effects of the searchers at the library, a fairly reasonable and reliable appraisal of what basic or core strategy gives substantial, if not 100%, recall of relevant citations can be obtained. And as Lancaster points out,

> When we consider that these ratios are merely tools by which we measure variations in performance within our own system, and within the confines of a controlled experiment, it is evident that any method that will give us reasonably accurate ~~estimates~~ of recall and precision is adequate, as long as we hold the method constant throughout the evaluation program. Even if the method results in slightly inflated, or slightly deflated, estimates of recall or precision, since the method is held constant it will still result in performance figures that will be valid tools to use in the comparison of system alterations (1968, p. 131).

Furthermore, both precision and recall must be used together in order to get an accurate picture of what occurs because of their inverse relationship, i.e., the more precise a search, the lower the recall and vice versa (Lancaster, 1968, pp. 56, 58-59).

At the heart of this study to determine effectiveness of search profiles is the concept of relevance. A source of longstanding, continuing discussion and debate, relevance has been considered in great depth (for two reviews on the subject see Rees and Saracevic, 1966 and Saracevic, 1975). One factor seems readily agreed upon, that relevance or similar evaluative judgments have definite subjective elements (cf. Swanson and Meyer, 1975, p. 143; Fugmann, 1973, p. 359; Saracevic, 1975; pp. 340, 341, 342; and Rees and Saracevic, 1966, pp. 9, 16). However, agreement also exists that relevance can be judged (Lancaster, 1968, pp. 120-121 and Rees and Saracevic, 1966, pp. 6, 10). "Although it may appear that relevance judgment is a very subjective human process, it has associated with it some remarkable regularity patterns" (Saracevic, 1975, p. 342).

Since relevance can be judged, we are led to the issue of who will judge and how the judgment will be made. This almost immediately raises the question of how relevance and pertinence differ. As Rees and Saracevic have pointed out,

> . . . a sharp distinction can be made between relevance to a question and relevance to the need underlying a question (documents satisfying the need are referred to by some authors as "pertinent", and the ones answering the question itself are "relevant": ,i.e., the real measure desired derives from the relation to the satisfaction of the information need of the user)(1966, pp. 8-9).

And, since

> Relevance is the property which assigns certain members of a file (e.g., documents) to the question; pertinence is the property which assigns them to the information need [,] . . . some relevant answers are also pertinent; but there could be relevant answers that are not pertinent and pertinent answers that are not relevant. It has often been argued that, from the user's point of view, desirable answers are pertinent aswers; but, in reality, an IR [information retrieval] system can only provide relevant answers. That is, a system can only answer questions. It can only guess what the information need is. In practice, there is often a real tug of war in trying to satisfy information needs and not just answer questions (Saracevic, 1975, p. 332).

Because pertinence is "the subjective assessment by a user against his own information needs" and "is confined to those aspects of an individual's situation that are of concern to him, and may change over time", its measurement tends to be very specific and individualized, whereas relevance "is capable of public assessment and can, therefore, only be assessed against a user's statement of his need" (Butterly, 1975, p. 190). Since health effects as a concept used in relationship to a chemical/pollutant is a broad category and is requested as a search topic by a wide variety of people from EPA researchers to administrators to state and local governmental agencies to private firms with government contracts, it is intended to give a wide view of a substance's generally accepted and potential health effects (usually negative but sometimes beneficial), encompassing the entire animal system. This coupled with

1) the tendency of differences in intended use of documents to produce differences in relevance judgments "suggesting that intended use becomes part of the query" (Saracevic, 1975, pp. 341-342) and these different user groups would have different intended uses; 2) that individual's preferences, purposes, and needs change leading to rejection of citations or similar citations that once satisfied or the reverse situation (Swanson and Meyer, 1975, p. 142); and 3) that the user is inclined to judge search responses to a request "with respect to his subjective, a priori undefinable information need . . . and not with respect to the objective, definable, and well considered search requirements" mainly because of the added time and concentration it would require to learn and conduct such analyses, "for this would divert them too much from their discipline-oriented activities" (Fugmann, 1973, pp. 361-362) would produce pertinence judgments and not the desired relevance judgments.

This plus the following factors led to the decision that the author, with the advice and assistance of the head librarian and searcher, should make the relevance judgments:

1) The difficulty in getting a representative sample of judges from the user population not only because of the searches supplied to non-EPA people but also because of the complexity of EPA organization.

2) Because of the author's and the head librarian's subject experience, by virtue of the area of searches performed and knowledge of the scope of health effects, they are better qualified to make this kind of relevance judgment than a panel of judges or an individual judge. Extensive, specialized subject knowledge is not necessary and would probably be a handicap.

3) Since there is no ranking of relevance order or evaluating whether partially or totally relevant, the relevance decisions are easier to make. The citations are judged either relevant or not. Nor is the quality of citations being evaluated because the primary purpose of this study is to determine the effectiveness of different search strategies in retrieving relevant citations.

Each citation was judged relevant if it dealt with any health effect whether negative or beneficial, although the emphasis is on negative effects, and whether of primary or secondary importance in the citation document. Excluded were documents that dealt only with diagnostic methods or treatment and did not discuss the actual health effects of asbestos.

All citations printed from each data base were displayed in the fullest format available for the data base to enhance relevance evaluation. Also, experimentation has shown that "the more complete a record, the more likely is its selection as a hit" (Schipma, 1976, p. 5).

The author first evaluated the citations for relevance, then the head librarian evaluated any citations whose relevance was in question. The author was responsible for making all final relevance judgments. Relevance judgments were then checked to ensure that the same relevance designation had been given to the same citation regardless of what subsection of a search or in what data base it appeared.

After relevance judgments were assigned, the relevant citations for each section were counted and the precision percentage calculated for the BIOSIS, Chemcon, and NTIS searches. In the Enviroline and Pollutiona Abstracts approach, the precision was calculated several times as a guide for profile development. More indepth analyses by manual methods followed and are discussed in the next chapter on data analysis. These methods were used to arrive at the final strategy(ies).

IV. DATA ANALYSIS

BIOSIS (See Figure 2 for strategy and data.)

In the BIOSIS strategies all codes selected are limited to major (primary and secondary) indexing to cut down on irrelevance. This designation of "major", when referring to codes in the text, is denoted by a preceding asterisk, e.g., *22506.

The essential strategy, Section 1, uses some toxicology codes, the carcinogen codes, and those for teratology. As can be seen in the composite, it exhibits high recall and acceptable precision of 68% (set 10). However, the asbestos strategy F87 creates irrelevancies in this file because the term serpentine (see Figure 1) sometimes refers to this soil type instead of to asbestos. This problem could be corrected by NOTing the following strategy against the final set in the asbestos serial:

1 SERPENTINE (3W) SOIL
2 SERPENTINE (3W) SOILS
3 SOIL (3W) SERPENTINE
4 SOILS (3W) SERPENTINE
5 C1-4/OR.

It would not, in practice, be worth the additional cost since relatively few citations are affected. If this correction is made by excluding irrelevant serpentine soil citations in the counts, the precision increases to 70%.

Section 2, the first nonessential strategy, tests pathology codes. All the citations overlap those of section 1. Set 18 contains 0 citations. The precision level is good, 83% (set 17).

Fig. 2. COMPOSITE OF BIOSIS SEARCH STRATEGY*

Section	Set	Postings	Search statement	Description	Result
	1	633	SERIAL# F87	(Asbestos stored strategy)	
	2	114	1/76000001-62070004	(1976 accession numbers)	
Section 1	3	24776	CC=22506	(Toxicology--Environmental)	
	4	3079	CC=22508	(--Veterinary)	
	5	32282	CC=24007	(Neoplasms and Neoplastic Agents)(Carcinogens and Carcinogenesis)	
	6	21030	CC=25552	(Teratology and Teratogenesis--Descriptive)	
	7	3440	CC=25554	(Experimental)	
	8	80090	3-7/OR		
	9	60393	8/MAJ		
	10	97	2AND9	(94 without Serpentine Soil)	66 Relevant (66) 68% Precision (70%)
Section 2	11	26787	CC=12502	(Pathology, General and Misc.--General)	
	12	21834	CC=12503	(--Comparative)	
	13	98093	CC=12504	(--Diagnostic)	
	14	18249	CC=38004	(Veterinary Science--Pathology)	
	15	148413	10-13/OR		
	16	73552	14/MAJ		
	17	12	2AND15		10 Relevant 83% Precision
	18	0	17NOT10		
Section 3	19	65295	CC=22501	(Toxicology--General, Methods, and Experimental)	
	20	10347	CC=37013	(Environmental Health--Occupational Health)	
	21	26607	CC=37015	(--Air, Water, Soil Pollution)	
	22	2958	CC=37019	(--Miscellaneous)	
	23	97402	19-23/OR		
	24	73145	23/MAJ		
	25	4086	Mutag?	(Mutagen(s), Mutagenic, Mutagenesis, etc.)	
	26	4669	Mutat?	(Mutate(s), Mutation(s), Mutating, etc.)	
	27	1869	Teratog?	(Teratogen(s), Teratogenesis, etc.)	
	28	401	Teratol?	(Teratology, Teratological, etc.)	
	29	17119	Carcinogen?	(Carcinogen(s), Carcinogenic, Carcinogenesis, etc.)	
	30	13547	Cancer?	(Cancer, Cancers, Cancerous, etc.)	
	31	24738	Tumor?	(Tumor, Tumors, etc.)	
	32	13581	Carcinoma?	(Carcinoma(s), etc.)	
	33	1507	Neoplasm?	(Neoplasm(s), etc.)	
	34	12777	SERIAL# CV1	(Mortality words stored strategy)	
	35	79044	25-34/OR		
	36	140643	35OR24		
	37	99	2AND36	(97 without Serpentine)	66 Relevant (66) 66% Precision (68%)
	38	7	37NOT10		2 Relevant 28% Precision
Section 4	39	22097	CC=12002	(Physiology, General and Misc.--General)	
	40	29343	CC=12003	(--Comparative)	
	41	43231	CC=22504	(Toxicology--Pharmacological)	
	42	91407	39-41/OR		
	43	55296	42/MAJ		
	44	5	2AND43		2 Relevant 40% Precision
	45	3	44NOT10		0 Relevant 0% Precision
Section 5	46	42074	CC=03506	(Genetics and Cytogenetics--Animal)	
	47	39581	CC=03508	(--Human)	
	48	82535	CC=13002	(Metabolism--General; Metabolic Pathways)	
	49	43376	CC=13003	(--Energy and Respiratory Metabolism)	
	50	42038	CC=13020	(--Metabolic Disorders)	
	51	213512	43-50/OR		
	52	158971	51/MAJ		
	53	7	2AND52		7 Relevant 100% Precision
	54	0	53NOT10		
Section 6	55	54393	CC=34502	(Immunology (Immunochemistry)--General; Methods)	
	56	6581	CC=34506	(--Immunohematology (includesBldGrps))	
	57	74844	CC=34508	(--Immunopathology (Tissue Immun.))	
	58	112119	55-57/OR		
	59	86507	58/MAJ		
	60	6	2AND59		4 Relevant 67% Precision
	61	3	60NOT10		1 Relevant 33% Precision

*This search was run in several sections because of storage overload problem for very large set sizes.

Section 3 tested three different things: 1) codes for general toxicology and environmental health; 2) free-text words for some concepts that are covered by CROSS codes (e.g., mutagenesis, carcinogenesis, etc.); and 3) the free-text words for mortality in Figure 3 that do not have specific CROSS code counterparts. While this section as a whole produced several

Fig. 3. Stored mortality words strategy CVI.

```
S MORTALIT?
S DEATH
S DEATHS
S FATAL?
S AUTOP?
S LETHAL
C 1-6/OR
```

more citations than Section 1 (cf. sets 10 and 27), all but seven overlapped those in Section 1 (set 38). Also Section 3's overall precision was 4% lower. The precision value of the seven unique citations was only 28%.

Section 4 tested general physiology codes and pharmacological toxicology. It yielded few citations (set 44), only 3 not included in the core strategy (set 45), and none were relevant.

Section 5 tested codes for genetics (i.e., mutagenesis) and metabolism in general. It produced no citations not included in the core (set 54) even though the citations it retrieved were all relevant (set 53). It is important to remember, however, that asbestos is not considered a mutagen _per se_ and therefore would not and did not retrieve enough citations in this area to make an evaluation.

Section 6 tested codes for immunology. Six citations were produced at 67% precision (set 60), but only 3 were not included in the core of which 1 was relevant (set 61) for a precision of 33%.

This so far has been a superficial and strictly numerical analysis of results. Manual delving into the individual sections produces a much more comprehensive and accurate picture of the strategy's effectiveness.

In Section 1 *22508 did not retrieve any citations. Veterinary toxicology was originally entered to see if it picked up relevant articles that discussed health effect of animals. This may not be significant. *25554 and *25552 did not appear and both are teratology codes, *25552 appearing in its only appearance together with *22506 and *24007. This, however, reflects the failure of asbestos to produce such effects and not the utility of the teratology codes in health effects search profiles. This will have to be tested later with other chemicals/pollutants as should *22508. Experience has shown these codes to be useful for many chemicals/pollutants.

This leaves *22506 and *24007 solely responsible for retrieval of all Section 1 citations, relevant and nonrelevant. Table 1 gives the actual numerical breakdown. As can be seen, both must be used since each

Table 1. Numerical breakdown of code appearance in Section 1

Section 1	Total Cits.	*22506	Alone	*24007	Alone	Together
Relevant Citations	66	64	22	44	2	42
Nonrelevant Citations	31	30	26	4	0	4

appears as the only index term from the core strategy in some of the citations. In checking the resultant citations from Section 2—which overlapped completely those of Section 1—*22506 would have retrieved them all and only *22506 and *24007 of the Section 1 strategy appear in these citations. The actual numerical breakdown appears in Table 2.

Table 2. Numerical breakdown of Section 1 code appearance in Section 2

Section 2	Total Cits.	*22506	Alone	*24007	Alone	Together
Relevant Citations	10	10	8	2	0	2
Nonrelevant Citations	2	2	2	0	0	0

The citations unique to the Section 3 strategy did not contain any of that section's keywords with the exception of one citation which was nonrelevant. The breakdown for the CROSS code appearances is in Table 3.

Table 3. Numerical breakdown of code appearance in Section 3

Section 3	Total Cits.	*22501	Alone	*37013	Alone	Together
Relevant Citations	2	1	1	1	1	0
Nonrelevant Citations	5	3	3	0	0	0

*37015 appeared once but only in a nonrelevant citation. *22501 and *37013 together would have retrieved all relevant unique section citations.

Since neither Section 4 nor Section 5 retrieved relevant citations different from those in Section 1, their codes need not be used as the core strategy picks up all of those retrieved. However, the genetics and cytogenetics codes *03506 and *03508 still need to be compared against the mutagenesis words with other substances to test the relative retrieval effectiveness of these two approaches. This is also true of the teratology codes in Section 1 versus the words in Section 3. The mortality words need to be checked against other chemicals/pollutants, especially those more immediately fatal than the slow-acting asbestos, to test their utility.

Section 6 retrieved two relevant citations unique from Section 1.

All three of the unique citations contained only *34508 from this section's strategy. However, both these relevant citations contain the two CROSS codes that retrieved all the Section 3 unique citations. The breakdown appears in Table 4. Since both of these relevant citations were included in the unique citations of Section 3, the Section 6 codes need not be used in the strategy.

Table 4. Numerical breakdown of Section 3 code appearance in Section 6

Section 6	Total Cits.	*22501	Alone	*37013	Alone	Together
Relevant Citations	2	1	1	1	1	0
Nonrelevant Citations	1	0	0	0	0	0

To arrive at the final strategy, then, only Sections 1 and 3 need to be considered for asbestos health effects. This is also probably true for health effects of other chemicals/pollutants. By combining the useful codes in these two sections, the following asbestos health effects strategy develops:

CC=22506 (Toxicology—Environmental)
CC=24007 (Neoplasm/Neoplastic Agents—Carcinogens/Carcinogenesis)
CC=22501 (Toxicology—General)
CC=37013 (Environmental Health—Occupational Health)
1-4/OR
5/MAJ

The results would be as follows:

97	(94 without Serpentine Soil)	Citations from Section 1 (Set 10)
7		Citations from Section 3 unique from Section 1 (Set 38)
104	(101 without Serpentine Soil)	Total Citations
66	Relevant Citations from Section 1	
2	Relevant Citations from Section 3 (Set 30) unique from Section 1	
68	Total Relevant Citations	

$\frac{68}{104} \times 100 = 65\%$ Precision or $\frac{68}{101}$ (correcting for Serpentine Soil) $\times 100 = 67\%$

This is still acceptably close to the desired 70% level.

Previous experience indicates that the overall search strategy for health effects in BIOSIS should include the two teratology codes in Section 1, even though testing with asbestos fails to bring this out. Thus, the strategy in Figure 4 is recommended.

Fig. 4. Recommended basic health effects strategy.

1 CC=22501
2 CC=22506
3 CC=22508
4 CC=24007
5 CC=25552
6 CC=25554
7 CC=37013
8 1-7/OR
9 8/MAJ

The author still recommends the testing of other parts of the strategy discussed on pages 23 - 25 of this paper.

Chemcon (See Figure 5 for strategy and data)

Section 1, the essential strategy for Chemcon, contains the combination frequently used by the library's searchers for health effects. It contains the general toxicology subject code and free-text and index terms. Health is limited to titles and descriptors to cut down on irrelevance. It produces a substantial level of recall on health effects of asbestos for this data base with a reasonably good precision of 74%

Section 2 is a group of free-text words and synonyms for cancer effects, mutagenesis, teratogenesis, and mortality. While it retrieved 10 citations at 100% precision, it added no new citations to the essential or core strategy. This is not surprising since the Chemcon subject code CA004 in Section 1 includes a subsection on chemicals including industrial chemicals and a subsection on carcinogens. However, as with the BIOSIS

Fig. 5. COMPOSITE OF C A CONDENSATES SEARCH STRATEGY*

Section	No.	Postings	Search statement	Description	Result
	1	2254	SERIAL# F87	(Asbestos stored strategy)	
	2	539	1/84000001-85999999	(1976 accession numbers)	
Section 2	3	38745	SC=CA004	(Toxicology)	
	4	17384	TOXIC?	(Toxic, Toxicity, Toxicology, etc.)	
	5	2467	HEALTH/TI,DE		
	6	1453	HYGIEN?	(Hygiene, Hygienic, etc.)	
	7	48827	4-7/OR		
	8	46	2AND8		3-4 Relevant 74% Precision
Section 2	9	3914	CARCINOGEN?	(Carcinogen(s), Carcinogenic, Carcinogenesis, etc.)	
	10	10138	CANCER?	(Cancer(s), Cancerous, etc.)	
	11	8328	TUMOR?	(Tumor(s) etc.)	
	12	1210	CARCINOMA?	(Carcinoma(s) etc.)	
	13	3801	NEOPLASM?	(Neoplasm(s) etc.)	
	14	2185	MUTAGEN?	(Mutagen(s), Mutagenic, Mutagenesis, etc.)	
	15	2750	MUTAT?	(Mutate(s), Mutations(s), Mutating, etc.)	
	16	1127	TERATOG?	(Teratogens, Teratogenic, Teratogenesis, etc.)	
	17	180	TERATOL?	(Teratology, Teratological, etc.)	
	18	1431	SERIAL# CVI	(Mortality words stored strategy)	
	19	27660	9-18/OR		
	20	10	2AND19		10 Relevant 100% Precision
	21	0	20NOT8		---
Section 3	22	24434	SC=CA059	(Air Pollution and Industrial Hygiene)	
	23	5387	SC=CA003005	(Biochemical Interactions—Mammalian Systems)	
	24	835	SC=CA003006	(—Human Systems)	
	25	237	SC=CA005005	(Agrochemical—Mammal (rodenticides, etc.))	
	26	9609	SC=CA013002	(Mammalian Biochemistry—Metabolism)	
	27	1336	SC=CA013004	(—Genetics)	
	28	13900	SC=CA013013	(—Other (gen. physiol. chem. stud.))	
	29	55722	22-28/OR		
	30	31	2AND29		5 Relevant 16% Precision
	31	24	30NOT8		1 Relevant 4% Precision
Section 4	32	3694	SC=CA014003	(Mammalian Pathological Biochem.—Metabol. & Hered. Diseases)	
	33	3678	SC=CA014004	(—Organ & Gland. Diseases)	
	34	762	SC=CA014005	(—Digest. & Excret. Diseases)	
	35	618	SC=CA014006	(—Reprod. Dis. & Preg.)	
	36	2062	SC=CA014007	(—Circul. & Resp. Diseases)	
	37	1077	SC=CA014008	(—Nervous & Sens. Diseases)	
	38	905	SC=CA014009	(—Blood Dyscrasis)	
	39	4777	SC=CA014010	(—Cancer (neoplasia))	
	40	1663	SC=CA014013	(—Other)	
	41	19236	32-41/OR		
	42	0	2AND41		— —

*Run as more than one search (two searches).

strategy, further tests on teratogenesis, mutagenesis, and mortality words need to be undertaken with other chemicals/pollutants.

Section 3 includes a variety of potentially useful subject codes from air pollution and industrial hygiene of the Applied Chemistry and Chemical Engineering Sections to biochemical interactions, mammalian biochemistry, and agrochemicals of the Biochemistry Sections. This produced a fair amount of citations but precision was a low 16%. Of the citations not included in Section 1, only one was relevant giving a 4% precision value.

Section 4 contains specific subsections of CAC14, the section on mammalian pathological biochemistry. It produced no citations on asbestos.

A more indepth analysis of Section 1 appears in Tables 5 and 6.

Table 5. Numerical breakdown of search term appearance in Section 1

Section 1	Total Cits.	SC= CACC4	Alone	Toxic?	Alone	HEALTH /TI,DE	Alone	HYGIEN?	Alone
Relevant Citations	34	30	22	10	0	5	2	0	—
Nonrelevant Citations	12	8	6	3	1	3	3	0	—

Table 6. Numerical breakdown of search term co-appearance in Section 1

Section 1	Total Cits.	SC=CACC4 & TOXIC? Together	TOXIC? & HEALTH/TI,DE Together	SC=CACC4, TOXIC?, & HEALTH/TI,DE Together
Relevant Citations	34	7	2	1
Nonrelevant Citations	12	2	0	0

None of the citations contained the free-text term HYGIEN?. SC=CACC4 was by far the most powerful in retrieving citations. TOXIC? retrieved no relevant citations alone, but HEALTH/TI, DE did.

Since Sections 2 and 4 added no additional citations to the Section

1 strategy, Section 3 remains for consideration. It added only one relevant citation (set 31) and had a low precision of ~~only~~ 4%. The only subject code in Section 3 retrieving citations was CA059. A breakdown of retrieval by its subsections is in Table 7. Constructing a strategy

Table 7. Numerical breakdown of CA059 subsection appearance in Section 3

Section 3	Total Cits.	000	Alone	001	Alone	002	Alone	003	Alone	004	Alone	005	Alone
Relevant Citations	1	0	–	0	–	1	1	0	–	0	–	0	–
Nonrelevant Citations	23	5	5	1	1	12	12	4	4	0	–	1	1

using search terms that retrieved relevant citations produces the following profile:

1 SC=CA004
2 HEALTH/TI,DE
3 SC=CA059002
4 1-3/OR

The results would be as follows:

46 Citations from Section 1
13 Citations using SC=CA059002
59 Total Citations

34 Relevant Citations from Section 1
1 Relevant Citation using SC=CA059002
35 Total Relevant Citations

$\frac{35}{59} \times 100 = 59\%$ Precision

Although CA059 includes toxicity of air pollutants to humans and other animals and industrial hygiene, particularly subsections 002 (Air pollutants and pollution) and 003 (Industrial hygiene), its scope is much broader than this and thus introduces a high level of irrelevance when used in a health effects search. A searcher might, with clear conscience, opt to leave it out, using the profile in Section 1. The author opts for the whole profile in Section 1 rather than the abbreviated "SC=CACC4 OR HEALTH/TI,DE" version of it that would retrieve all relevant citations for asbestos health effects but the SC=CAC59CC2 indexed one. TOXIC? and HYGIEN? need to be tested with other chemicals/pollutants, especially with ones whose effects are more immediate than those of asbestos. The words portion of the strategy also warrants further investigation.

NTIS (See Figure 6 for strategy and data.)

When dealing with the subject codes preceded by CF=, one digit codes must be entered both with and without the preceding C placeholder, e.g., CF=C6T? and CF=6T?. This is because these codes have been applied both ways at varying times. All codes are truncated because "*s" indicating use as a major descriptor have also been variously applied.

Section 1 of the NTIS search contained HEALTH/TI,DE,ID plus the index codes for Environmental Health; Environmental Biology; Industrial Medicine; Public health, hygiene, and industrial medicine; and Toxicology. Only 2C citations, a fair amount, were retrieved as the date base is smaller than BIOSIS and Chemcon (set 17). However, its precision of 65% is good for this data base since NTIS tends to have a high level of irrelevancy because of indexing and abstracting practices.

Section 2 contained the mortality words strategy CVI (Figure 3) which produced no citations about asbestos (set 19).

Section 3 contained index codes for pathology, genetics, physiology, and chemical and biological warfare which also produced no citations about asbestos (set 27).

Section 4 contains free-text words on carcinogenesis, mutagenesis, and teratogenesis. This section had low retrieval and mediocre precision

Figure 6. NTIS SEARCH STRATEGY

Section	#	Postings	Statement	Description
	1	340	SERIAL# F87	(ASBESTOS stored strategy)
	2	2	1/A7185H2-A729214	(1976 accession numbers)
	3	52	1/C5571C1-C780514	
	4	1	1/D0011A1-D0115F4	
	5	55	2-4/OR	
Section 1	6	882	CF=68G?	(Environmental Health)
	7	1020	CF=06F?	(Biol. and Med. Sciences—Environmental Biology)
	8	3689	CF=6F?	(" " " " " ")
	9	315	CF=06J?	(—Industrial (occup.) medicine)
	10	756	CF=6J?	(" " ")
	11	619	CF=06T?	(—Toxicology)
	12	2433	CF=6T?	(")
	13	1606	CF=57U?	(—Public health, hygiene, & ind. med.)
	14	2000	CF=57Y?	(—Toxicology)
	15	11503	HEALTH/TI,DE,ID	
	16	18661	6-15/OR	
	17	20	5AND16	13 Relevant 65% Precision
Section 2	18	2709	SERIAL# CVI	(Mortality words stored strategy)
	19	0	5AND18	— —
Section 3	20	840	CF=57O?	(—Pathology)
	21	1397	CF=57E?	(—Cytology, genetics, & mole. biol.)
	22	294	CF=06P?	(—Physiology)
	23	5651	CF=6P?	(—Physiology)
	24	1698	CF=57S?	(—Physiology)
	25	301	CF=74D?	(Military Sciences—Chemical, biological, and radiol. warfare)
	26	8367	20-25/OR	
	27	0	5AND26	— —
Section 4	28	506	CARCINOGEN?	(Carcinogen(s), Carcinogenic, Carcinogenesis, etc.)
	29	1261	CANCER?	(Cancer(s), Cancerous, etc.)
	30	347	TUMOR?	(Tumor(s), etc.)
	31	112	CARCINOMA?	(Carcinoma(s), etc.)
	32	1560	NEOPLASM?	(Neoplasm(s), etc.)
	33	316	MUTAGEN?	(Mutagen(s), Mutagenic, Mutagenesis, etc.)
	34	807	MUTAT?	(Mutate(d), Mutating, Mutations(s), etc.)
	35	25	TARATOG?	(Teratogen(s), Teratogenic, Teratogenesis, etc.)
	36	122	TERATOL?	(Teratology, Teratological, etc.)
	37	3589	28-36/OR	
	38	4	5AND37	2 Relevant 50% Precision
	39	1	38NOT17	0 Relevant 0% Precision

of 50% (set 38). It added no new relevant citations to the Section 1 strategy.

By analyzing Section 1 in greater depth, two breakdowns of code appearance are possible and are shown in Tables 8 and 9. For health

Table 8. Breakdown of search term appearance in Section 1

Section 1	Total Cits.	CF=68G?	Alone	CF=C6F? or 6F?	Alone	CF=C6J? or 6J?	Alone	CF=C6T? or 6T?	Alone	CF=57U?	Alone	CF=57Y?	Alone	HEALTH/TI,DE,ID	Alone
Relevant Citations	13	8	0	0	–	10	0	10	0	9	0	4	0	10	1
Nonrelevant Citations	8	3	0	3	1	1	0	1	0	2	0	0	–	5	3

Table 9. Use of Section 1 search terms by relevant and nonrelevant citations

	Cit. No.	CF= 68G?	CF= C6F? or 6F?	CF= C6J? or 6J?	CF= C6T? or 6T?	CF= 57U?	CF= 57Y?	HEALTH /TI,DE,ID
Relevant Citations	1				X		X	
	2			X		X		X
	3	X		X		X		X
	4							X
	5				X		X	X
	6	X		X	X	X	X	X
	7	X		X	X	X		X
	8	X		X	X	X		X
	9	X		X	X	X		X
	10			X	X			
	11	X		X	X	X	X	X
	12	X		X	X	X		X
	13	X		X	X	X		X
Nonrelevant Citations	1							X
	2	X	X					
	3							X
	4		X					
	5							X
	6	X		X	X	X		X
	7	X	X			X		X

effects of asbestos, only two search terms would have been needed, HEALTH/TI,DE,ID and CF=6T? or C6T? The precision level also would have increased to 72% since two nonrelevant citations (numbers 2 and 4 in Table 9) would not have been retrieved.

However, it would be premature to generalize so spartan a strategy from the results for asbestos. Notice, for example in Tables 8 and 9, that all but CF=C6F?/CF=6F? appeared in the relevant citations. Based on this core strategy, the author believes that it should remain as is but be tested further for verification, especially CF=C6F?/CF=6F?. TOXIC? was not used in this test and might be worth testing against the toxicology codes. As with BIOSIS and Chemcon, the words and/or codes on carcinogenesis, mutagenesis, teratogenesis, and mortality need further testing by using other chemicals/pollutants.

Enviroline

By selecting all the citations on the asbestos terms for the volumes corresponding to 1975 through 1976 (see Figure 7), it was possible to work backwards to arrive at a strategy. First, relevant citations were

Figure 7. Asbestos search strategy for years 1975 through 1976.

1	235 SERIAL# F87 (Asbestos stored strategy)
2	97 1/800000-1199999 (1975-1976 Accession numbers)
	36 Relevant 37% Precision

determined. Then likely candidates for search terms were circled in all the citations. The circle terms in nonrelevant citations were later used to calculate total citations that would be retrieved from a strategy and its precision.

The number of relevant articles containing each term were counted.

HEALTH, HAZARD?, and HYGIEN? were further subdivided by whether they appeared in the title and/or descriptor of the citations or in any other location of the citation. See Table 10.

Table 10. Breakdown of tentative search term appearance in relevant citations

Tentative Search Term	No. of Citations
RC=02 (Chemical and Biological Contamination)	31
HEALTH/TI,DE	25
HEALTH not in TI,DE	15
HAZARD?/TI,DE	2
HAZARD? not in TI,DE	7
TOXIC?	2
CARCINO?	21
CANCER?	21
PATHOL?	18
DISEASE?	8
DISORDERS	5
EXPOSURE?	15
ADVERSE	2
HYGIEN? in TI and/or DE	2
HYGIEN? not in TI,DE	3
BIOLOGICAL	5
EPIDEMI?	1
MORTAL?	1
DEATH	3
FATAL?	1
NEOPLASM?	2
MALIGNAN?	1

A tentative group of words was selected and evaluated as in Table 11. RC=02 is not included in this table.

Table 11. Appearance of tentative search terms in relevant and non-relevant citations

Terms	No. of Relevant Citations	No. of Appearances Alone	No. of Nonrelevant Citations	No. of Appearances Alone
HEALTH/TI,DE	25	0	17	5
HEALTH/AB	13	1	12	2
HAZARD?	7	1	5	3
CARCINO?	21	0	16	5
CANCER?	21	0	11	1
PATHOL?	18	0	4	1
EXPOSURE	15	0	8	0
DEATH	3	0	0	0
TOXIC?	2	0	3	1
TOTAL CITS.	36 Relevant		37 Nonrelevant	

Since only a very small number of the search words that were considered important, either because of actual numbers of citations retrieved or relationship to health effects terminology (e.g., TOXIC?), appeared as the only tentative search term in a citation; each citation was checked for each of the terms in Table 11 plus RC=O2. See Table 12.

Table 12. Appearance of tentative search terms in relevant citations

Cit. No.	HEALTH /TI,DE,AB	HAZARD?	CARCINO?	CANCER?	PATHOL?	EXPOSURE	DEATH	TOXIC?	RC+ O2
1	X								X
2		X							X
3	X		X	X		X	X		X
4	X				X	X			
5	X				X				
6	X		X		X				
7	X								X
8			X		X				X
9	X		X	X					X
10	X				X	X			X
11	X		X			X			
12	X				X				X
13	X					X			X
14	X	X	X	X					X
15	X								X
16	X		X	X	X				X
17	X	X		X		X			X
18	X		X	X					X
19				X	X				X
20	X		X	X					X
21		X	X	X	X				X
22	X	X	X	X					X
23			X	X	X				X
24	X		X	X	X	X	X		X
25	X		X	X	X	X	X		X
26	X			X	X	X			X
27	X		X	X		X			
28	X		X	X		X			X
29	X			X		X			X
30			X	X					X
31			X	X	X				X
32			X	X	X				X
33	X	X		X	X	X			X
34	X		X		X				
35	X	X				X		X	X
36	X		X	X	X	X			X

Table 13 shows the eight citations which are not retrieved by HEALTH/TI,DE,AB. By using the health effects strategy in Figure 8

Table 13. Appearance of search terms in the eight citations not retrieved by HEALTH/TI,DE,AB

Cit. No.	HAZARD?	CARCINO?	CANCER?	PATHOL?	EXPOSURE	DEATH	TOXIC?	RC= C2
2	X							X
8		X		X				X
19				X			X	X
21	X	X	X	X				X
23		X	X	X				X
30		X	X					X
31		X	X	X				X
32		X	X	X				X

with F87 (see Figure 1), all 36 relevant citations would be retrieved.

Figure 8. Enviroline health effects strategy number 1.

1 HEALTH/TI,DE,AB
2 HAZARD?
3 PATHOL?
4 CARCINO?
5 1-4/OR

Looking at the nonrelevant citations using these search terms and RC=C2 in Table 14, shows 37 nonrelevant citations would be retrieved. Figure 9 shows the statistics for the resultant search strategy.

Figure 9. Effectiveness of strategy number 1.

36 Relevant Citations
37 Nonrelevant Citations
73 Citations

$\frac{36}{73}$ X 100 = 49% Precision

This precision is an important improvement over just selecting asbestos, but still modest.

Table 14. Appearance of tentative search terms in nonrelevant citations

Cit. No.	HEALTH /TI,DE,AB	HAZARD?	CARCINO?	CANCER?	PATHOL?	EXPOSURE	DEATH	TOXIC?	RC=02
1									X
2	X		X		X	X			X
3					X				X
4	X					X			X
5	X	X							
6			X						
7	X								
8									X
9	X					X		X	X
10									X
11									X
12			X						X
13		X							X
14			X						X
15									X
16								X	
17									X
18	X	X							X
19	X								X
20									X
21	X					X			X
22			X	X					X
23			X						X
24									X
25		X							X
26									X
27	X		X			X			X
28	X								
29	X								X
30	X								
31		X							
32			X	X					X
33	X								
34									X
35			X	X					X
36	X		X	X		X			X
37			X						X
38	X			X					X
39	X		X	X		X			X
40	X			X		X			X
41				X					X
42				X	X				X
43	X			X					X
44	X		X						X
45	X		X	X					X
46			X		X				X
47			X					X	

Another simpler search approach would be the following:

1 HEALTH/TI,DE,AB
2 RC=02
3 1OR2

This would result in

36 Relevant Citations
43 Nonrelevant Citations
79 Total Citations

$\frac{36}{79}$ X 100 = 47% Precision

While this lowers the precision, it has the advantage that RC=02 (Chemical and Biological Contamination) includes the numerous health-related Enviroline keywords within it, e.g., carcinogenic agents; health, env; pathology, human. This still needs to be tested with other chemicals/pollutants, but the approach looks favorable. This level of precision is certainly acceptable for this data base because of the way in which it is indexed.

It would also be wise to test the longer strategy further, especially in the work areas of toxicology, mortality, mutagenesis, teratogenesis, and carcinogenesis.

Pollution Abstracts

The approach for developing the health effects strategy in this data base was very similar to that for Enviroline. Pollution Abstracts is even more specific in its indexing than Enviroline and has may fewer indexing words from which to choose. Also, Pollution Abstracts has no indexing codes.

All citations for the asbestos serial #F87 were selected and then manually limited to the years 1972 - 1976. This could also be done in the following way:

1 SERIAL# F87
2 YR=76
3 YR=75
4 YR=74
5 YR=73
6 YR=72
7 2-6/OR
8 1AND7

The result was 95 citations, 48 of which were relevant, with a precision of 51%.

After the relevant citations were determined, likely candidates for search terms were circled in all citations. The number of relevant articles containing each term were counted and whether or not this candidate was the only candidate for search term appearing in the citation was noted. Table 15 shows this approximate count. Then ten terms were selected

Table 15. Appearance of tentative search terms in relevant citations

Tentative Search Term	Number of Relevant Citations	Number of Citations Where Search Term Appeared Alone
HEALTH/TI,DE	21	2
HEALTH not in TI,DE	12	1
HYGIEN?	4	2
HAZARD?	6	1
PATHOL?	14	3
DISEASE?	9	2
TOXIC?	3	1
CYTOTOXIC?	4	2
CARCINO?	12	1
CANCER?	4	0
MUTAG?	1	0
TERATO?	1	0
EPIDEMI?	3	1
MORTAL?	1	0
AUTOP?	1	1
BIOLOGIC?	4	0
No terms	2	-

to be checked for appearances in relevant citations on the basis of their being the only candidate search term used in at least one relevant citation. Table 16 gives this listing.

Table 16. Appearance of tentative search terms in relevant citations

Cit. No.	HEALTH /TI,DE	HEALTH not in TI,DE	HAZARD?	CARCINO?	PATHOL?	DISEASE?	HYGIEN?	CYTO-TOXIC?	EPI-DEMI?	AUTOP?
1					X	X				
2	X	X	X							
3	X	X			X					
4	X	X		X						
5	X								X	
6	X		X		X					
7	X				X					
8	X	X		X						
9	X									
10		X		X	X					
11	X			X						
12	X	X					X			
13	X					X				
14	X		X							
15				X	X					
16				X	X					
17	X								X	
18					X					
19				X	X					
20				X						
21					X	X				
22					X					
23	X									
24	X			X						
25					X					
26	X	X			X					
27	X	X			X	X				
28				X		X				
29						X				
30		X								
31		X		X						
32		X		X			X			
34										X
35								X		
36								X		
37								X		
38						X				
39	X					X				
40	X		X			X				
41										
42	X		X							
43									X	
44							X			
45							X			
46			X							
47	X	X						X		
48										

Note that after selecting the 10 terms for the tentative strategy, the frequency of some of the terms being the unique search term increased.

These ten terms would retrieve all but two of the relevant citations. In this case, the citations lost are some of the less specifically health effects citations. One is on the fringe and concerns enzymes, asbestos, and detergents; the other concerns the amount of chrysotile asbestos in lungs of New York City residents, or bioaccumulation. The latter falls into the area of a specific syndrome, i.e., effects on lungs, that would be picked up in a search specifically interested in the pulmonary health effects of asbestos. But the purpose of this study is more generally oriented, as stated earlier, and thus the general strategy of the ten search terms fills the health effects needs.

Checking for these ten search terms in the nonrelevant citations produced the results in Table 17.

Table 17. Appearance of tentative search terms in nonrelevant citations

Cit. No.	HEALTH /TI,DE	HEALTH not in TI,DE	HAZARD?	CARCINO?	PATHOL?	DISEASE?	HYGIEN?	CYTO-TOXIC?	EPI-DEMI?	AUTOP?
1	X									
2		X		X			X			
3			X							
4	X									
5	X						X			
6			X							
7	X									
8				X						
9		X								
10		X								
11	X									
12	X	X								
13	X		X				X			
14					X	X	X		X	
15	X									
16		X								
17		X					X			

The strategy would be

1 HEALTH
2 HYGIEN?
3 CARCINO?
4 HAZARD?
5 EPIDEMI?
6 DISEASE?
7 PATHOL?
8 CYTOTOXIC?
9 AUTOP?
10 1-9/OR

Normally in searching health effects, HEALTH is limited to TI and DE and sometimes also ID and AB. However, in Pollution Abstracts, HEALTH appears as the only search term outside these boundaries and, thus, must not be so limited. Consolidating the 2 "HEALTH"s shortens the strategy to nine terms.

The result of this strategy would be as follows:

46 Relevant Citations
17 Nonrelevant Citations
63 Total Citations

$\frac{46}{63}$ X 100 = 73% Precision

Several terms need more testing. These are the toxicity, mutagenesis, teratogenesis, and mortality words. Since asbestos is not a substance which rapidly produces toxic effects, it does not adequately test this concept or that of mortality.

V. CONCLUSION

In running these tests again, the author suggests truncating AMPHIBOLE to AMPHIBOLE? in the asbestos serial to ensure picking up citations under the group name amphiboles in the rare event when asbestos might not appear in the citation or abstract. Also, the term MALIGNAN? should be added to the portion of strategies testing cancer words. Its absence in this case is not critical. After developing the strategies for Enviroline and Pollution Abstracts, it appears that the terms EXPOS? and HAZARD? should also have been tested in Chemcon and NTIS.

As stressed throughout this study, the strategies and results are tentative. All five data bases need testing in several areas, including toxicology, mutagenesis, carcinogenesis, teratogenesis, and mortality words. Different substances which produce such effects need to be tested to see how useful these subdivisions of searches are. By ascertaining how useful these elements are in retrieving relevant citations for substances known to produce these effects, a sort of checklist for health effects of substances can be produced. For example, certain codes in a given data base have been verified as retrieving, at an acceptable level of precision, relevant citations on these subdivisions of health effects when the free-text words in this area do not. If these same codes are then used in a search on a particular health effect, such as teratogenesis, and no relevant citations are produced, the probability is high that no documents/articles on this effect have been entered into the data base.

Although this is a preliminary study, definite trends still can be ascertained for these data bases by testing with asbestos, and are represented in the results for each data base. Codes work well in BIOSIS; Chemcon and NTIS require a combination of codes and words; on Enviroline either words alone or words and codes together can be used; and Pollution Abstracts, of course, requires all words whose identities can be pinpointed as was done in this study.

APPENDIX A. LIST OF AIDS IN PROFILE DEVELOPMENT

*BioSciences Information Service of Biological Abstracts. BIOSIS Search Guide: BIOSIS Previews Edition. Philadelphia, Pa.: BIOSIS, 1977.

_____. CROSS Code. Philadelphia, Pa.: BIOSIS, n.d.

_____. A Guide to the Vocabulary of Biological Literature. Philadelphia, Pa.: BIOSIS, 1973.

_____. Profile Guide. Philadelphia, Pa.: BIOSIS, n.d.

_____. Subject Guide to the CROSS Index. Philadelphia, Pa.: BIOSIS, n.d.

Chemical Abstracts Service. Subject Coverage and Arrangement of Abstracts by Sections in Chemical Abstracts. 1975 edition. Columbus, Ohio: American Chemical Society, 1974.

Environment Information Center, Inc. Enviroline User Manual. New York: Environment Information Center, n.d.

Lockheed Information Systems. Brief Guide to DIALOG Searching. Palo Alto, Ca.: Lockheed Information Services, 1976.

Lockheed Retrieval Services Information Systems Laboratory. DIALOG Terminal Users Reference Manual. 2 vols. Palo Alto, Ca.: Lockheed Missiles and Space Co., n.d.

Lockheed Information Systems. Online information on data bases and limiting. March 13, 1977.

Stamm, Roy P., and Ryerson, Ted. NTIS Subject Classification (Past and Present). Springfield, Va.: National Technical Information Service, Nov. 1975. NTIS/SR-75101.

Strategy cards prepared by EPA-RTP library's searchers.

Using CA Condensates and CASIA. Presented at the DIALOG User's Workshop, Chicago, Ill. July 16-17, 1976.

Pollution Abstracts. Pollution Abstracts Keyword Master List. Louisville, Ky.: Data Courier, 1976.

*Received at EPA-RTP library after BIOSIS search strategies were developed and run.

BIBLIOGRAPHY

Becker, F. H.; Veal, D. C.; and Wyatt, B. K. Comparative efficiency of searching titles, abstracts, and index terms in a free-text data base. Journal of Documentation 28:22-36 March 1972.

Beauchamp. R. O., Jr.; Daugherty, M. A.; Garber, J. L.; and Myers, J. D. Comparative searching of computer data bases. Journal of Chemical Documentation 13(1):32-35 1973.

Bogovski, P.; Timbrell, V.; Gilson, J. C.; and Wagner, J. C. (eds.). Biological Effects of Asbestos: Proceedings of a Working Conference. Switzerland: International Agency for Research on Cancer, World Health Organization, 1973. IARC Scientific Publication No. 8.

Butterly, E. Improving SDI search profiles. Information Processing and Management 11(8-12):189-200 1975.

Byrne, J. R. Relative effectiveness of titles, abstracts, and subject headings for machine retrieval from the COMPENDEX services. Journal of the American Society for Information Science 26:223-9 July 1975.

Calkins, M. L. On-line services and operational costs. Special Libraries 68(1):13-17 January 1977.

Callaghan, P. M. and Howden, M. E. H. Education in the use of computerized chemical information systems. Australian Academic and Research Libraries 3:207-14 December 1972.

Durkin, K. and Smith, J. R. Retrieval forms and formats for environmental objectives employing machine-readable biological information. Journal of Chemical Information and Computer Sciences 16(1):14-16 1976.

Farradane, J. and Yates-Mercer, P. A. Retrieval characteristics of the index to Metal Abstracts. Journal of Documentation 29:295-314 September 1973.

Fisher, H. L. and Elcheson, D. R. Effectiveness of combining title words and index terms in machine retrieval searches. Nature 238:109-10 July 14, 1972.

Fugmann, R. On the role of subjectivity in establishing, using, operating and evaluating information retrieval systems. Treatise II on retrieval system theory. Information Storage and Retrieval 9:353-372 1973.

Heim, K. M. The role of mechanized services in the provision of information with special reference to the university environment. Masters thesis, University of Wisconsin, 1975. ERIC document ED-119 725.

Hines, D. L. Computerized literature search services in an engineering library. Special Libraries 66:197-204 April 1975.

Hunter, L. J. Selective dissemination of information in the environmental sciences: computerized data bases available for this service. Masters research paper, University of North Carolina at Chapel Hill, April 1975.

Jenkins, G. T. The MEDLARS demand search quality control program. Bulletin of Medical Library Association 60(3):423-6 July 1972.

Lancaster, F. W. Information Retrieval Systems: Characteristics, Testing and Evaluation. New York: Wiley, 1968.

Lancaster, F. W. and Fayen, E. G. Information Retrieval On-Line. Los Angeles: Melville Publishing Co., 1973.

Lancaster, F. W.; Rapport, R. L.; and Penry, J. K. Evaluating the effectiveness of an on-line natural language retrieval system. Information Storage and Retrieval 8(5):223-45 1972.

Laurence, B.; Weil, B. H.; and Graham, M. H. Making on-line search available in an industrial research environment. Journal of the American Society for Information Science 25:364-9 Nov./Dec. 1974.

Long, C. W. Characteristics of the bibliographic retrieval of on-line data bases: a comparison of BIOSIS Previews, Chemical Abstracts Condensates, Excerpta Medica, MEDLINE, and TOXLINE. Masters research paper, University of North Carolina at Chapel Hill, 1976.

Lynch, M. F. Computer-based Information Services in Science and Technology: Principles and Techniques. Stevenage, England: Peregrinus, 1974.

Martin, T. H. Reflections upon the state-of-the-art in interactive information retrieval. In Information Systems and Networks. Ed. by John Sherrod. Westport, Conn.: Greenwood Press, 1975.

McCullough, C. B. Computerized data bases in the field of environmental health: MEDLINE, TOXLINE, SCISEARCH, POLLUTION, and CAIN compared. Masters research paper, University of North Carolina at Chapel Hill, 1975.

Nees, M. and Green, H. O. The BIOSIS data base: evaluation of its indexes and the STRATBLDR, CHEMFILE, STAIRS, and DIALOG systems for on-line searching. Research Triangle Park, N.C.: N.C. Science and Technology Research Center, 1976.

Prewitt, B. G. On-line searching of computer data bases. Journal of Chemical Documentation 14(3):115-7 1974.

Prewitt, B. G. Searching the Chemical Abstracts Condensates data base via two on-line systems. Journal of Chemical Information and Computer Sciences 15(3):177-83 1975.

Rees, A. M. and Saracevic, T. The measurability of relevance. Comparative Systems Laboratory, Technical Report No. 7. Cleveland, Ohio: Western Reserve Univ., 1966.

Roe, K. E.; Micuda, V.; and Seeds, R. S. Literature searching with the CAIN on-line bibliographic data base. Bioscience 25(12):796-7 1975.

Saracevic, T. Comparative effects of titles, abstracts, and full texts on relevance judgements. In Proceedings of the American Society for Information Science, vol. 6. Ed. by J. B. North. Westport, Conn.: Greenwood, 1969.

Saracevic, T. Relevance: a review of and a framework for the thinking on the notion of information science. Journal of the American Society for Information Science 26:321-43 Nov./Dec. 1975.

Saracevic, T. and Rees, A. M. Towards the identification and control of variables in information retrieval experimentation. Comparative Systems Laboratory, Technical Report No. 6. Cleveland, Ohio: Western Reserve Univ., 1966.

Scheffler, F.; March, J.; and Bernados, J. An experiment to study the use of Boolean not logic to improve the precision of selective dissemination of information. Journal of the American Society for Information Science 23:58-65 Jan./Feb. 1972.

Schipma, P. B. Computer search center statistics on users and data bases. Journal of Chemical Documentation 14(1):25-9 1974.

Schipma, P. B. Research toward enhancing retrieval effectiveness. Washington, D.C.: National Science Foundation, 1976. ERIC document ED-121 302.

Schultz, L. Breaking the communication barrier between searcher and literature file: an interactive guide. Journal of the American Society for Information Science 25:3-9 January 1974.

Shearer, J. An appraisal of the RECON system as operated at the Science Reference Library. Geoscience Documentation 7(5-6):98-100 December 1975.

Smith, S. W. Venn diagramming for on-line searching. Special Libraries 67(11):510-17 November 1976.

Standen, A. (ed.). Kirk-Othmer Encyclopedia of Chemical Technology, 2nd ed. vol. 2 "Asbestos," pp. 734-747. New York: Interscience, 1963, 1967.

Stevens, M. E. Strategies for organizing and searching. In National Information Retrieval Colloquium, 1973, Changing patterns information retrieval. Philadelphia: ASIS, 1974. Pp. 47-79.

Swanson, R. W. and Mayer, J. Performing evaluation studies in information science. Journal of the American Society for Information Science 26:140-56 May/June 1975.

Verheijen-Voogd, C. and Mathijsen, A. A contribution to the comparison of the usefulness of the data bases of Excerpta Medica and MEDLARS in biomedical literature retrieval. ASLIB Proceedings 26:136-151 April 1974.

Weiss, I. Evaluation of ORBIT and DIALOG using six data bases. Special Libraries 67(12):574-581 December 1976.

Williams, M. E. Use of machine-readable data bases. In Annual Review of Information Science and Technology, vol. 9, 1974. Ed. by Carlos A. Cuadra and Ann W. Luke. Washington, D.C.: ASIS, 1974. Pp. 221-84.