thousands of words annually. Most of them, of course, are chemical names. The incidence of low-frequency new words in CA is very high, and any inverted file design has to allow for many additions each year.

Finally, just to prove that one can find respite from all these statistics, note the peculiar set of terms with similar frequencies (Figure 17) that appeared on the computer printout one day when we were running a frequency-ordered term frequency distribution list on *CA Condensates.*

LITERATURE CITED

(1) Williams, M. E. Schipma, P. B. Preece, S. E., Becker, D. S., Llewellen, P. A., and Stewart, A. K., "Educational and Commercial Utilization of a Chemical Information Center," IIT Research Institute, Computer Search Center, Chicago, Ill., June 25, 1968–June 25, 1972; NSF Contract No. **NSF-0554,** July 30, 1972; available NTIS or ERIC Clearinghouse, **ED 068 132,** hardcover price $16.45.

Profiling, the Key to Successful Information Retrieval*

C. H. O'DONOHUE
Box 26583, Philip Morris Research Center, Richmond, Va. 23261

Received October 9, 1973

A major tool employed to enter an information source is the search profile. The development of an adequate profile depends upon the aids supplied by the data bases. These aids vary in their content and depth and their proper use is essential for relevant information retrieval. The data bases examined are *CA Condensates, Index Medicus,* and BA data bases. Several searches are presented with a study of their comparative profiles.

The useful yield of any literature search will depend upon the requestor's ability to communicate effectively with the information system (data base). The usual communication tool used by the requestor is known as the profile or search question.

DEVELOPING THE PROFILE

Needs. The requestor has certain particular needs which must be met. The search strategy (Figure 1) begins when the requestor contacts a profiler or becomes a profiler himself. As part of this search strategy, an understanding of the structure of the indexing language used in the data base is required. The requestor/profiler has no control over the input into the system nor does he have any control over the method of input. He must operate within the constraints placed on him by the data base.

He needs to know what keywords (descriptors, indexing terms, codens) are available for the input phase and how these terms are defined and related. The profile structuring process is very similar to that of document indexing. Thus, before a requestor can begin, he is dependent on the data base for the items needed to help him structure his profile. In most cases he will have to subordinate his wishes to the idiosyncrasies of each data base he encounters.

What help do the various commercially available data bases give to a potential user? If one subscribes to magnetic tape services, then a word-frequency listing can be obtained. Also data bases supply certain aids to their magnetic tape subscribers that are not available to the hard copy (published edition) subscribers. Depending upon whether one uses outside commercial services or does manual searches in-house, the profile structuring problems are quite different.

Profiling Aids. The three major data bases that will be discussed are *Biological Abstracts* (BA), *Chemical Abstracts* (CA), and *Index Medicus.* The indexing/profiling aids available from these services are varied in depth of content.

Biological Abstracts provides a "Guide to the Indexes for Biological Abstracts and Bio-Research Index" and "A Guide to the Vocabulary of Biological Literature." The Guide to the Indexes only illustrates the "how-to-use" approach to the four indexes of *Biological Abstracts* and *Bio-Research Index.* The second, the Guide to the Vocabulary, is of major value in that it points one to the right road for choosing the proper keywords.

B.A.S.I.C., the subject index for BA, is a permuted, keyword-in-context index to published titles. These titles contain augmenting keywords added by BA. The Guide points out the pitfalls involved in the B.A.S.I.C. indexing concept, as for example, in a study involving dogs, such keywords as dog, dogs, canine, puppy, and beagle would have to be checked to obtain the relevant citations. The Guide contains word frequency listing and related term listing and was designed to aid those using either the published indexes or the magnetic tape data base.

As might be expected, *Chemical Abstracts* has many aids available for the user. Subscribers to the complete CA also receive the "Index Guide," "Index to Ring Systems," "Formula Index," and the "Registry Number Index." Also available are the "CAS Search Guide," "The Desktop Analysis Tool (DAT)" and a "CAS Chemical Substance Name Selection Manual for the Ninth Collective Index." Generally, each of these aids is needed at one time or another. Figure 2 lists two dyes chosen from DAT. Fewer than 6% of the possible names for each of these substances appeared in the "Index Guides." Thus, a compound with a common name having neither a registry

*Presented before the Division of Chemical Literature, 166th Meeting, ACS, Chicago, Ill., Aug. 28, 1973.

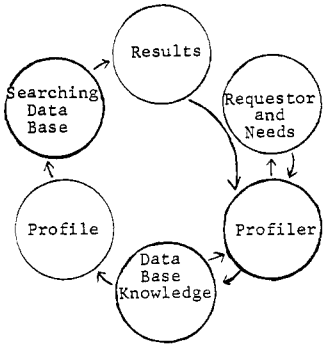

Figure 1. Search strategy

Syndrome/Tumor (MeSH Term)	X-Referenced in MeSH
Guillain-Barré (Polyradiculitis)	Yes
Werner-His Disease (Trench Fever)	No
Brooke's Tumor (Skin Neoplasms)	No
Ondine's Curse (Hypoventilation Diabetes Insipidus Appetite Disorders Hypothalamus Brain Disease)	No

Figure 3. MeSH terms

number nor formula would be difficult to identify without DAT or a similar tool. At present DAT covers only compounds of biomedical interest. As with most tools, it should be used with caution, and the obtained name and registry number verified in another source, as some of the earlier numbers listed in DAT have since been deleted. The "CAS Search Guide" is intended primarily for users of the magnetic tape services, but with proper use it can aid in manual searches.

Index Medicus has a rigidly controlled vocabulary. The authority list for this data base is "Medical Subject Headings" (MeSH). It is useful to a limited degree. Better aids are the "Medlars Indexing Manual" and the "Medlars Indexing and Searching Aids." These indexing tools can be of great value to the profile developer. For example, while some syndromes and tumors are listed in MeSH, others do not appear (Figure 3). The last three terms are found only in the "Indexing and Searching Aids." These two manuals also indicate which terms, if not present in MeSH, are used in MEDLINE, the magnetic tape data base.

It has been stated that searchers need to be trained for indexing in order to structure profiles properly.[1,2] The fact that the indexing aids are often of the greatest value tends to support this statement.

USING THE PROFILE

Manual. Manual searching refers to a search conducted in-house using a published data base. The usual steps in developing an adequate profile are (1) select the data base, (2) select the keywords, (3) run a sample search, and (4) modify the profile. After a request has been made and the searcher and requester have discussed the question, the appropriate data base is chosen. After the data base has been selected, the search profile is developed. The first profile developed is used to search a representative portion of the data base. The search results are examined to determine if the retrieved material is relevant. If the search question is not satisfied, then the profile may have to be modified. The modified profile is then used for the entire data base unless changes in indexing procedures during various time periods necessitate additional changes in the profile.

Computer (Magnetic Tape). All magnetic tape searches needed by the Philip Morris R&D Center are handled by outside service centers. The problems of obtaining an adequate profile under these circumstances are:

A. Misunderstanding of question by service center
B. Service center personnel not knowledgeable in certain data bases
C. Selected keywords are too general
D. Selected keywords are too specific

The advantages of an outside service center handling the profile development are:

A. Personnel more proficient in the use of the data base
B. Term frequency listing readily available
C. Authority lists for numeric codes available

Service centers are also aware of many of the problems previously cited.[1]

Some service centers attempted to have the users do their own profile coding. As Carmon has pointed out,[2] this procedure fails because the user usually doesn't have access to such aids as word frequency lists, authority lists for numeric codes, and the magnetic tape thesauri. Also, the investment in staff necessary to train the user is as great as when the staff develops the profiles for the users. Thus, there is no cost benefit for the service center.

CA Name	Registry No.	Number of Names DAT	Number of Names Index Guide
C. I. Pigment Green 7	1328-53-6	82	5
C. I. Leuco Sulphur Blue 7	1327-57-7	89	1

Figure 2. Dyes

EVALUATING THE PROFILE

Some case histories (Figure 4) illustrate some of the problems experienced in structuring and implementing an adequate profile.

Case No. 1

Problem: The design of a specified type of oven or drying tunnel

Data base: *Engineering Index*

Profile: The profile was formulated by an outside service center and not revealed

Response: There were 51 abstracts of which only two dealt with design. The rest dealt with the applications of such ovens

Evaluation: The question was misunderstood by the service center. Upon further questioning it was found that this question had been combined with another question which the center thought was similar so that only one profile had to be run

Case No. 2

Problem: The pyrolysis structure of carbohydrates

Data base: *Chemical Abstracts*

Case No.	In-House	Service Center	Profile Revealed	Percent Abstracts Relevant*
1	No	Yes	No	4
2	Yes	Yes	Yes	95
3	Yes	Yes	No	27
4	Yes	Yes	Yes	54

* All in-house searches are considered 100% relevant. This column applies only to outside searches.

Figure 4. Profile cases

Table I. Condensed Profile for Case No. 4

Group	Term No.	Term	Group	Term No.	Term
G01	1	Ammonia	G05	10	Dissociate*
G01	2	HN_3	↓	↓	↓
G02	3	Decompose*	G05	15	Degrade*
G02	4	Decompose*			
G02	5	Cracked	G06	16	Heat*
		Ammonia	G06	17	Thermal*
			↓	↓	↓
G03	6	Nitrogen	G06	31	Discharge*
G03	7	N_2			
			G06	32	Gas
G04	8	Hydrogen			
G04	9	H_2			

Boolean Logic is G01*(G02 + (G03*G04) + (G05*G06)).

Profile: The in-house profile required 62 major keywords to describe the carbohydrates. Each term was associated with 10 minor keywords. The service center required a total of only 36 keywords

Response: There was 95% relevancy with no misses for the service center

Evaluation: Through the use of truncation and authority lists the service center was able to obtain all relevant abstracts with a minimal number of descriptors. The use of Boolean logic and truncation permitted the use of 23 terms to describe the carbohydrates plus 13 terms to describe the possible pyrolysis processes

Case No. 3

Problem: Formation of a flavor material by pyrolysis

Data base: *Chemical Abstracts*

Profile: The flavor material's chemical name was used as the keyword, and entries to its preparation were checked. The service center in its search strategy also used one reference which we were aware of prior to our search

Response: The service center found only borderline citation

Evaluation: There was no way in which a profile could be adequately constructed to search the data base. The term pyrolysis could not be used as entries under this term referred to "pyrolysis of." Only by searching under the chemical name and examining every entry which indicated the compound was formed in some manner, could completeness be ensured

Case No. 4

Problem: The reaction of ammonia during its decomposition

Data base: *Chemical Abstracts*

Profile: The in-house search was conducted by scanning all entries under the term, "ammonia." The service center required a total of 32 terms

Response: The service center cited 67 abstracts with a relevancy of 54%

Evaluation: By examining all subterms under the major heading of ammonia, the in-house manual searcher ensured completeness of the search. This required more time than if a profile had been constructed. The service center's profile also ensured completeness but gave false retrieval of items dealing with decomposition reactions occurring in NH_3, the formation of NH_3 during a degradation process, and NH_3 degrading other materials (see Table I)

CONCLUSIONS

A data base or information source is entered through the use of a tool known as a profile. The desired information can be obtained only by properly structuring this profile.

Adequately trained personnel supplied with appropriate aids are essential for the correct structuring of a profile. A desirable background can be obtained by receiving the training of an indexer for the data base. This permits the acquisition of an in-depth knowledge of keywords and their usage.

Among the more important aids for profiling are the manuals that are used by the indexers. Word frequency lists and authority lists are also essential.

In general, an outside search on a magnetic tape data base by an adequately trained and staffed service center can be of great value in the completeness of the findings and in time savings. However, unless the service center can be relied on to produce a satisfactory profile, manual searches will remain the better way because the requestor can modify his keywords and profile or even change data base while conducting the search.

LITERATURE CITED

(1) Park, M. K., *Spec. Lib.* **64,** 187-92 (1973).
(2) Carmon, J. L., *Spec. Lib.* **64,** 65-9 (1973).