ACCESS PROBLEM

Access is an interesting problem in and of itself. Access refers to the ability of an IAC scientist or engineer who has need of some information to get it—when he needs it, as he needs, with a minimum of muss, fuss, bother, and bureaucratic nonsense. The relevance recall ratio should be excellent, otherwise he is likely to complain about irrelevant material. And, if he is forced to go to many places to obtain the coverage he demands, he will voice a strong objection.

The IACs I'm familiar with suffer little from access problems. This is due to the intimate and continued participation of the user in the collection, storage and retrieval, and using phases. However, when the IAC information operation utilizes more general collections—i.e., relatively broad information stores—access can be a severe problem. As seen through our IAC eyes, the access problem of the general collection is caused by the idiosyncracies of indexing; specifically, an indexing operation designed for other than our specific areas of technology.

The major access problems faced by the types of IACs we operate stem from two conditions—that access problem caused by proprietary interests, and that access problem caused by national defense interests. Both of these interests are understandable and justified. No ready solution to these problems has been offered.

End of Symposium

Amortization of Indexing (Input) Costs*

JULIUS FROME
U. S. Patent Office, Dept. of Commerce,
Washington, D. C. 20231

Received September 3, 1971

A formula is given that enables one to determine the parameters for amortizing an information system within a year. The formula is illustrated with three mechanized information systems for patents: (1) mold and mold coating composition class 106, subclass 38.2–38.9; (2) synthetic fibers class 264–210, 288, 289, 290; and (3) abrasives class 51, subclass 298–309. The most critical parameter is shown to be indexing time.

The cost of information systems resolves itself into several main factors:

1. System design
2. Machine costs
3. Indexing (input)
4. Retrieval (output)

System design as a one time proposition is a minimal cost in the over-all picture. Machine cost is also not of too great importance since in many instances computers and punch card equipment are already present for other purposes or if not the acquisition of these will lead to so many other uses that the cost of the machines would be a small part of the over-all cost of the system. Retrieval costs are minimal. The major cost of an information system is the indexing of the file, particularly in the time of the scientist or professional. Clearly, the quality of search by a well designed and indexed machine retrieval system is certainly good or better than a manual search.[1-3] This paper describes a technique of lowering the cost of indexing and possibly amortizing the main cost of indexing in one year.

THE PROBLEM

Most organizations operate on an annual budget. It would be desirable, therefore, if the cost of the information system could be amortized in the period of one year. That is, the benefit could accrue in the same period of time as the cost of the system (one year). As indicated supra, the main cost of an information system would be in the time used in indexing. Thus, if the time saved in using a system for search would be equal to or greater than the time spent in indexing, the cost could be amortized in one year. This would be very important for management to be able to plan effectively within the annual budget.

CRITERIA OF SECTION OF ART

To achieve our purpose, the art must have the following criteria:

1. The art must be sufficiently active
2. The manual search must take a fair amount of time
3. The important or commonly used terms must be able to be developed
4. The amount of time saved per search times the number of searches needed to be made in one year should be equal to or greater than time needed to index each patent times the number of patents—i.e.,

$$T \times N_s = I \times N_D$$

T = time saved per search (machine over manual)
N_s = number of searches per year
I = indexing time per document
N_D = number of documents in system

This formula can be used for example as follows:

Let us say there are 2000 documents in an art. The time saved per use of the mechanized system is about six hours. Let us say that there are about 50 searches in a year. Therefore in accordance with the formula:

*Presented at 3rd Northeast Regional Meeting, ACS, Buffalo, N. Y., Oct. 11, 1971.

$$TN_s = IN_D$$
$$6 \times 60 \times 50 = I \times 2000$$
$$I = 9 \text{ minutes}$$

Therefore, to amortize the search in one year, you can allow nine minutes of indexing time per document. Of course, if there are 100 searches a year you can allow 18 minutes per document or if there were only 1000 documents you can allow 18 minutes per document.

THE ART SELECTED

The following three were selected from Class 106, and subclasses 38.2–38.9, (Molds & Mold Compositions); Class 51, subclasses 298, 307–309 (Abrasives); and Class 264, subclasses 210, 288–290 (Synthetic Fibers). The definitions of these classes and subclasses can be found in the "Manual of Classification."[7]

The first two groups are composition classes—i.e., mixtures of materials.

The third group, although it will handle compositions in certain forms, is mainly a process class.

It was estimated that the mold art had about 2000 patents, the Abrasives had about 1800 patents, and the synthetic Fibers had about 1000 patents. It was estimated that saving in time per patent application searched for the mold was six hours, the time for abrasive was six hours, and for synthetic fibers about eight hours. By an estimate of number of searches to be made in a year, we could obtain the time which should not be exceeded for indexing each patent. It appeared that the three above arts meet the criteria set forth supra.

METHODOLOGY IN SELECTION OF TERMS

The user in any information and retrieval system is probably the most important person. The main function of an IR system is to serve the customer. Too little thought has been given to obtention and retention of user confidence. In these systems, the user was consulted about the system, trained enough in the techniques of information retrieval, and participated in all aspects of the system. A very thorough examination of needs for a system was made. A thorough analysis was made of the documents as a whole entity.

In patents, the claims indicate what sort of questions had been asked in the past. We were fortunate that the users were experienced in their art. Therefore, they could estimate the kind of questions which might be asked in the future. Since we started with the premise that we would like to amortize the system in a year, this meant that time for indexing of the terms from the document would be limited. We made a definite decision that the system would not try to answer all possible questions put to it, but be able to search for the vast majority of instances. It has long been realized that a certain amount of effort might handle 75 to 85% of a piece of work, but it might take three times that effort to handle the other 15% of the work. Therefore, the systems were designed to handle most of the searches needed for examination, but there would be searches not possible to be handled by the system and would have to be handled manually. Since our time of indexing was limited, we decided that about 300 very frequently used terms should be sufficient. As this had been previously found to be very efficient in these other areas,[1, 4] we decided to use a printed code sheet containing these terms.[1]

These terms were arranged so that a chemist could easily use them. It was realized that although about 300 terms would give us a great many possibilities, it might restrict us in the future. Also, there was a great deal of information present in the document that would not be extracted. Some way had to be found to extract such information quickly and easily and render the system completely open ended. For those documents which had terms or concepts which were important to the document, the indexer merely wrote those terms on the side or bottom of the code sheet in red ink. This method took very little extra time and allowed the indexer to use as many terms beside those printed on the code sheet. These terms are eventually put on separate punch cards and sorted and alphabetized and put into a printed list as will be discussed infra. (See also "Semi-Automatic Indexing" by Frome.[5])

DETAILS OF CODES

Mold Code Sheets. The mold code sheets contained nine major groups:

1. Binder
   14 sub groups
2. Aggregate
   6 sub groups
3. Metals (cations)
   80 sub groups
4. Anions
   20 sub groups
5. Gases
   5 sub groups
6. Organic
   30 sub groups
7. Surfactants
   33 sub groups
8. Inorganic-Acids
   8 sub groups
9. Miscellaneous
   20 sub groups

In other words, the nine major groups were subdivided into 215 groups under them. The arrangement provided for genus and species and were designed for ease of finding the term. The code merely has to be circled in red. The code sheet is 10½ × 16 inches. Provision is made for a nine-digit document number and patent classification.

Abrasive Code Sheet. The abrasive code sheet had 19 major subdivisions:

1. Resins
   96 sub groups
2. Metals
   26 sub groups
3. Oxides
   25 sub groups
4. Mis (abrasives)
   6 sub groups
5. Flourides
   3 sub groups
6. Sulfides
   3 sub groups
7. Carbonates
   3 sub groups
8. Borides
   4 sub groups
9. Carbides
   11 sub groups
10. Silicates
    17 sub groups
11. Grains
    9 sub groups
12. Structure
    7 sub groups
13. Metals
    80 sub groups
    Anions
    30 sub groups
14. Adhesive
    8 sub groups
15. Substrate
    17 sub groups
    Materials
    25 sub groups
16. Processes
    7 sub groups
17. Wetting agent
    6 sub groups
18. Organic
    38 groups
19. Nitrides
    2 sub groups

The 19 major subdivisions were divided into 423 sub groups. The code sheet is 10½ × 16 inches and arranged by genus and species. Provision is made for nine-digit document number as well as patent office classification.

Synthetic Fibers. The synthetic fibers code sheet contained 15 major groups:

1. Resins
   96 sub groups
2. Processes
   133 sub groups
3. Bicomponent
   20 sub groups
4. Nozzle geometry
   6 sub groups
5. Mis
   14 sub groups
6. Dimension (parameters)
   27 sub groups
7. Draw ratio
   14 sub groups
8. Temperatures
   20 sub groups
9. Per cent shrink
   9 sub groups
10. Mis
    4 sub groups
11. Metal (cations)
    14 sub groups
12. Anions
    20 sub groups
13. Organic
    38 sub groups
14. Treating media
    11 sub groups
15. Functional media
    21 sub groups

The major groups are divided into 433 sub groups. The code sheets, $10\frac{1}{2} \times 16$ inches, have provision for a nine-digit document number and patent office classification.

STATISTICS OF INDEXING

Time of Indexing. The following times are for these specific code sheets and these specific arts with certain specific trained examiners. The average examiner indexing had many years experience in the area.

Molds: 5.1 minutes per patent
Abrasives: 5.8 minutes per patent
Synthetic Fibers: 11.5 minutes per patent

The mold composition and abrasive are mainly composition, and the synthetic fibers are mainly process.

Amortization. The following is calculated from the formula:

$$T \times N_s = I \times N_D$$

The time saved per patent application searched averaged about six hours in each art.

Thus, for the molds:

$$6 \times 60 \times N = 5.1 \times 1987$$
$$N_s = 28 \text{ searches}$$

Thus, 28 searches are needed to amortize the indexing time.

In the Abrasive Art:

$$6 \times 60 \times N_s = 5.8 \times 1818$$
$$N_s = 29.2 \text{ searches}$$

In the Synthetic Fiber Art:

$$6 \times 60 \times N_s = 11.2 \times 1108$$
$$N_s = 34.4 \text{ searches}$$

In actual fact, many more searches were made, in the above areas than needed for amortization in one year.

SEARCH

Searching is by coordination. By asking for the least frequent term first, a considerable time was saved in the computer search.

To assist the examiner in his search, a frequency chart of the code sheet was prepared. The following statistics were noted.

Mold

273 number of terms recorded
27861 frequency count

The recording of terms range from 0 to 1372 (binders).

Abrasive

421 terms recorded
30202 frequency count

The range is from 0 to 952 (oxides).

Synthetic Fibers

444 terms recorded
46218 frequency count

The recording of the terms range from 0 to 1064 (Resins).

The number of patents in the various arts were:

Mold art	1987
Abrasives	1818
Synthetic Fibers	1108

The average number of terms per patent was:

Mold	14
Abrasives	17
Synthetic Fibers	42

The computer search was performed on a Honeywell 1200 computer with 65 K and 6 tape drives.

The program allowed 12 searches per pass. The average search took 1 minute.

The searches were also performed on a multi-column sorter (Census Bureau) which sorts about 500 cards per minute or varying from 2 to 4 minutes per search. (See also "Punched Cards *vs.* Random Access Computer" by Frome.[6])

However, in addition to the computer or multicolumn search, the examiner has an alphabetical list of the terms with the patent number next to the term for those less frequently occurring terms. Thus, in a desired search, if the term or terms are on the printed code sheets, then the computer or multicolumn sorter is used. When the term is not on the printed sheet, then the alphabetical list is used. Since the alphabetical list was of the less frequent terms, there is usually very few documents containing the terms used. These printed alphabetical lists are kept at the examiner's desk and result in very fast searches where applicable.

The statistics for the alphabetical list are as follows:

Molds	3459 entries
Abrasives	4952 entries
Synthetic Fibers	2120 entries

SEARCH RESULTS

The following are some of the statistics in some of our searches. The quality of our searches were excellent.

Mold	No. of doc.	No. of searches
In 119 searches	0–5	63
	6–10	16
	11–20	14
	20–40	12
	40+	14
		119 searches

Of course, those situations where there were too many documents retrieved, additional descriptors might be added to reduce the number of documents per search.

Abrasives	No. of doc.	No. of searches
In 128 searches	0–5	64
	6–10	14
	11–20	22
	20–40	16
	40+	12
		128 searches

Synthetic Fibers	No. of doc.	No. of searches
Total 231 searches	0–5	102
	6–10	31
	11–20	43
	20–40	35
	40+	20
		231 searches

In many instances one mechanized search per patent application was sufficient. However, in other instances more than one search per patent application was used.

CONCLUSION

The above described technique has resulted in amortizing the indexing costs in one year.

LITERATURE CITED

(1) Frome, Julius, "A Punched Card System for Searching Steroids," *U. S. Patent Office Research and Development,* Rept. No. 7, 1957.

(2) Frome, Julius, "A Punch Card System for Phosphorus Compounds," *J. Chem. Doc.* 1, 84-7 (1961).

(3) Frome, Julius, "Random Access Mechanization of Phosphorus," *J. Chem. Doc.* 1, 76 (1961).

(4) Frome, Julius, and O'Day, Paul T., "A General Chemical Compound Code Sheet Format," *J. Chem. Doc.* 4, 33-42 (1964).

(5) Frome, Julius, "Semi-Automatic Indexing and Encoding," *U. S. Patent Office Research and Development,* Rept. No. 17, 1959.

(6) Frome, Julius, "Mechanized Searching of Phosphorus Compounds III Punched Cards *vs.* Random Access Computer," *J. Chem. Doc.* 1, 88-90 (1961).

(7) U. S. Dept. of Commerce, Patent Office *Manual of Classification,* Washington, D. C., July 1971.

Evaluation of the Database *CA Condensates* Compared with *Chemical Titles*

INGE BERG HANSEN

The Documentation Department, The National Technological Library of Denmark (DTB), Lyngby, Denmark

Received October 20, 1971

The performance of *CA Condensates* and *Chemical Titles* based on analysis of precision and "relative recall CT/CC" for a collection of 46 search profiles was studied over a period of one year. Special emphasis was laid on the function of the keyword phrases of CC and the users' attitude towards literature categories not represented in CT. The results are discussed in terms of the value of the systems for Danish users seen from the users' and the documentalist's point of view.

When Chemical Abstracts Service in September 1968 made their *CA Condensates* tapes available for use by national documentation centers and private companies all over the world, The National Technological Library of Denmark (DTB) decided to make a study of the new database to see if the user community in Denmark would respond favorably to the idea of having current access by computer to the world's chemical literature.

DTB had in 1968 been running the *Chemical Titles* tapes (CT) in cooperation with I/S DATACENTRALEN on commercial terms for the Danish scientific and industrial community for two years and had, therefore, developed a considerable amount of know-how in connection with computer-based information retrieval.[5-8,10]

The *Chemical Titles* service had from the start been run on a cost-recovery basis in the sense that the subscribers paid the actual computer costs, whereas the library paid the subscription charges to Chemical Abstracts Service. When the library decided to take up the more comprehensive and more expensive *CA Condensates* service, we were aware that the experiment might show that it might not be feasible to run the service on an ecomonically sound basis in a small industrial and scientific community like the Danish.

PURPOSE

Our primary intention in the present study was to compare *CA Condensates* and *Chemical Titles* and, further, to