Table V
Comparison of Time with Observed Cell Entries for the Retrieval Profile of 184 Search Queries in the Organometallics File of 3,625 Documents

	Indexing Procedure		
	Single indexer[a]	Double indexer[a]	Single indexer reviewed[a]
Indexing time (min.)	64.3	128.6	111.6
Total retrieval $(\bar{x}_{1\cdot})$	22.7	34.5	26.7
False retrieval $(\bar{x}_{12})$	4.5	9.2	5.6
Correct retrieval $(\bar{x}_{11})$	18.2	25.3	21.1
Missed documents $(\bar{x}_{21})$	9.6	2.5	6.7
Total relevant $(\bar{x}_{\cdot 1})$	27.8	27.8	27.8

[a] These cell entries are averaged over all queries. The cell entries in Table IV are for queries involving 4, 8, and 12 terms only.

acceptable level of indexing accuracy. The indexing quality control program (see Figure 1) is established to ensure that an acceptable level of indexing accuracy is maintained. The QC procedure involves testing indexing accuracy of periodic samples. Quality control acceptance tables can be constructed for various acceptance levels of p_3 and sample sizes. One may wish to test batches of documents prepared over periods of time or by a particular group. Furthermore, one may use these tests to establish a learning curve for future reference.

A more complete discussion of quality control techniques applied to indexing problems may be found elsewhere.[6] Quality control of indexing is particularly important in coordinate index systems since the retrieval errors are highly sensitive to indexing errors.

One must keep in mind that the evaluation procedures described provide specific pieces of information which should be added to a fund of knowledge concerning the entire information retrieval system. Modifications of the search system or retrieval system may be suggested as a result of information found in the evaluations performed during the file development. Need for further training of the indexers may be evident as a result of the indexing experiment. Perhaps the rules of indexing may need modification before indexing is continued. It is even possible that the original system may be abandoned and another approach formulated. The searchers may be instructed to broaden their search queries to reduce retrieval errors attributable to indexing errors.

Thus, the decisions which are made as a result of evaluation during file development can touch every aspect of system operation. It is important that the information upon which these decisions can be based be made available as quickly as possible.

[6] D. W. King, and J. M. Daley, *ADI Proceedings*, 1, 389 (1964).

A Central Information Retrieval System*

D. L. ARMSTRONG and M. T. GRENIER
Aerojet-General Corporation, Azusa, California

Received December 4, 1964

INTRODUCTION

This paper describes a centralized technical information retrieval system for a highly technically oriented company having over 30,000 employees of whom 6,000 are engineers. At the time of the establishment of the system, seven locations were involved. Six were separated by distances ranging from 5 to 400 miles, while the seventh was on the opposite coast.

Simply stated, the problem confronting us was the design and the establishment of an efficient, economical information retrieval system which would permit the maximum utilization of technical information generated at each of the several company plants. It is the purpose of this paper to present our chosen solution to this problem and, in so doing, to emphasize the reason for each decision made.

The several plants of the company are widely separated, and the subject matter includes a broad scope of science and technology. Although each of the several locations has a specific area of responsibility, there are many areas of mutual interest. Exchange of information among these locations did not always occur easily and in a systematic manner even within one plant. Finally, there was no single repository for all technical information. Consequently, the Corporate Technical Information Center (CTIC) was established with the mission of providing a comprehensive index to all company-generated technical information regardless of point of origin and ensuring that new information was brought to the attention of appropriate personnel.

* Presented before the Division of Chemical Literature, 145th National Meeting of the American Chemical Society, New York, N. Y.

BACKGROUND

The company was established in 1942 and grew at an extremely rapid rate since 1946. Early collections of documents were small, but each plant was provided at its start with an independent library. As a result, the indexing systems in the several libraries were not necessarily compatible, and each library had a collection of documents not necessarily duplicated elsewhere. At various times and in various locations, indexing of small collections of documents was accomplished with conventional library indexing and with the use of edge-notched cards, uniterm cards and dual dictionaries, and computers. The retrieval of technical information was an individual plant responsibility. The experience gained through the use of all these systems was thoroughly considered in designing the new system which was to serve the entire corporation. Other corporations and agencies having similar collections of documents and similar problems were visited to examine first-hand how they handled their problem. Finally, the pertinent literature was reviewed and assessed for procedures and philosophies directly pertinent to our situation.

PROGRAM ACTION

New documents of lasting value are currently being produced within the company at a rate estimated to be 2,500 per year. The system selected had to absorb these documents and answer the questions generated by 6,000 engineers doing research and development. The first step taken was to ensure that copies of all documents to be put into the system were sent to one location. To accomplish this, a corporate directive was issued which described desirable documents as being "those of permanent value or probable future interest." The decision of the importance of the documents was, in general, left to the heads of the departments issuing the documents. Formal technical reports and proposals published through normal channels were secured by placing the CTIC on the automatic distribution lists.

The indexing system adopted was the coordinate index. The selection of a coordinate index system meant that many different terms could be coordinated for specific answers, but that the time involved in providing a means for this correlation could be short. The reports we process usually are assigned an average of 40 terms each, a number almost impossible to handle on cards which must be typed or filed individually. Our previous experience with coordinate indexing had proved that it was capable of the depth of indexing desired. The questions posed by our technical people are usually very specific and can best be satisfied by detailed answers. The questions can sometimes be unanswerable; *e.g.*, "What kinds of bearing have operated in a space environment for two years without attention?"

A hierarchical system of indexing was much too inflexible for the variations in information it was necessary to handle. Thus, a single report might deal at significant length on the subjects of propellant formulations, strengths of materials, fabrication techniques, missile guidance problems, animal toxicity tests, and qualifications of newly assigned personnel.

The choice between manual and machine correlation of index terms was debated at length. Prominent among the considerations were frequency of questions, speed of response, and cost of input and retrieval. The frequency of inquiries ranged from one to perhaps twenty per day. Thus, it was perfectly feasible to employ a manual search procedure. At the present time inquiries average 12 per day.

Invariably, the answers to specific questions are needed immediately. This requirement ruled out the possible use of computers which would have incurred an overnight delay, inasmuch as the Center could not afford the expense of a computer sitting idly by during a major portion of the day, nor could it be accorded a priority which would permit it to interrupt the accounting or technical work assigned to our rather extensive installation of computers. We also found that the time required for the conversion of a technical question into computer language was greater than the actual search time needed for the system finally adopted. The additional time required for the search by the computer made its use even less desirable.

The selection of the means of establishing correlation among descriptors was next considered. Previous experience with dual-dictionaries of tabulated numbers and with optical coincidence cards led quickly to the choice of the latter as being much faster and more flexible with regard to altering search terms and excluding unwanted terms. Optical coincidence cards are available in a variety of shapes and forms; the number of documents to be put into the system (initial input estimated at 10,000) excluded the small capacity cards. It should be noted in passing, however, that we have seen examples of cards having a capacity of 500 report addresses applied to systems containing over 13,000 documents. The very large cards, those having a capacity of 40,000 addresses, were not chosen because of their unwieldy size. These considerations led to the selection of the 10,000 address Termatrex card, which measures 9¾ × 11½ in. This particular card system has the added advantage that automatic reading equipment is available which can convert the card information to a computer readable form should the number of reports of current interest ever exceed the capacity of the cards as extended by an easily manipulated number of frames of a microfilm strip.

The Termatrex cards are available in an alphabetically filed and randomly filed form. Although the virtues of the random form were not fully appreciated at the time, this variation was chosen. After 2½ years of experience, the authors can firmly state that this was the correct choice. The random cards eliminate the occurrence of cards "lost" by incorrect alphabetical filing. They minimize the need for any "see" references in the Thesaurus and permit the extensive inclusion of synonyms without difficulty. Finally, the speed of card location and of refiling is much greater than with alphabetically filed cards.

The answer to a question (*i.e.*, the response form handed to the engineer) consists of a single page containing a complete identification of the document, the abstract, and the descriptors used to index the document. The inclusion of the descriptors on the abstract form provides an auxiliary type of abstract and permits the rapid identification of false drops should any have occurred. What appear to be bound terms among the descriptors may appear in the Termatrex card file as separate words. They are grouped on the form to indicate clearly their association in the report. This is particularly applicable to chemical terms. A loose alphabetical arrangement of the descriptors on the form is used to avoid accidental repetition of descriptors and to save clerical time in posting the terms.

The Thesaurus is the key to the entire operation. It is built from words appearing in the documents themselves rather than from a previously designed Thesaurus. Words not appearing in the documents are added, of course, when appropriate, in order to index the material adequately.

Multiple correlations of synonyms were established within the Thesaurus. An example of the kind of synonym relationship established would be the four words: position, location, station, and fix, each of which appeared in separate reports, but all having the common meaning of geographic position expressed in terms of latitude and longitude. All four terms were assigned to the same Termatrex card. Thus, regardless of which term was used in framing the question, all documents concerning this concept would be recovered. The purpose, of course, was to lessen the burden on clerical personnel and to minimize the possibility of a missed document. The use of "see also" references was kept to a minimum and generally was employed only when a word was not truly a synonym with another already in the system. The problem of placing two synonyms in the Thesaurus without establishing an internal correlation by assigning both terms to the same Termatrex card was minimized by placing entire control of the Thesaurus in the hands of one person. This has remained the case, except that the indexers consult with each other on the feasibility of the inclusion of new words and the final decision is made by one person.

A special problem was presented by the numerous references to propellant formulations, each of which is numbered. These numbers now cover several thousand formulations. To avoid assigning one card to each formulation, the numbers were assigned in groups of 10 formulations to one card.** Thus, formulations 2531 through 2540 are all assigned to a single card. The possibility of false drops thus incurred was fully recognized, but this has not yet proved to be a problem. Because only a relatively small portion of the reports involve names of chemical compounds, and again to conserve cards, these chemical compounds are split apart except for the most important and most frequently recurring ones. This is in agreement with the practice recently instituted by ASTIA (now DDC) in its Chemical Thesaurus dated December 1962.

Consideration was given to the use of roles, but assignment of a role to a single chemical compound in the face of the multiple roles the compounds frequently play in one report resulted in confusion and an excessive utilization of cards. For this reason roles were not used.

A computer is used to update the Thesaurus. A Hollerith card is prepared for each new descriptor; revised computer print-outs are made periodically. A copy of each Thesaurus is sent to each plant library at the time of its release and furnished to each indexer. Advantage was taken of the fact that the descriptor terms were on computer cards by having an inverted index printed, which permitted the collection of all synonyms and the detection of possible omissions or unintentional duplications in the assignment of descriptor terms to Termatrex cards.

PERSONNEL

The manager of the CTIC was selected from the technical staff, on the premise that he could request and obtain cooperation from the various plants—a very necessary facet, we felt, in the success of this corporate venture.

Indexers having extensive scientific backgrounds were selected initially in the belief that such backgrounds were necessary to recognize the absence of critical words in the report which should be included among the descriptors in order to ensure retrieval of that document. In practice, this theory worked rather well; however, the indexers being more inclined toward the experimental side of the work rather than the literature, found the repetitive nature of the work distasteful, with the result that excessive employee turnover took place. A change in policy then occurred, which resulted in the employment of persons with technical backgrounds and experience as indexers.

An important factor in the operation of the Center includes weekly meetings of all personnel concerned, at which problems developed during the course of the previous week's work are discussed and resolved. Also, at regular intervals, company-produced motion pictures dealing with a single phase of our work are shown, as available, to the entire group so that the words with which they are working are given a pictorial reality. Attention is also given to the industrial engineering aspects of the work in order to minimize the clerical effort and to eliminate duplication of effort. Thus, the indexers work with a single abstract and identification form from which a reproducible master is prepared. No other auxiliary pieces of paper are employed. The Termatrex card files are mounted on a round lazy susan for ready access.

MECHANICS

Information retrieval is conducted through the various plant libraries. This is done for two reasons: (1) the very practical reason of permitting the engineer to obtain all pertinent information from one source convenient to his work station, and (2) to continue to use the libraries as the central source of all technical information and incidentally avoid any action which could be interpreted as an intrusion by the Corporate Technical Information Center in an area previously allotted to the libraries. Duplicate

** An improvement on this system has subsequently come to our attention. It consists of assigning one of each of the terms 0-9 to one of ten cards, one of each of the terms 10-90 to one of ten other cards, and so on through 1000-9000. It would thus be possible by drilling the cards reading "4000," "200," "60," and "3," to pinpoint the number "4263." Forty cards are sufficient to locate reports dealing specifically with any one of 9999 propellant formulations. The same concept may be extended (although less precisely) to authors' initials through the use of 26 cards assigned to the letters of the alphabet. Greater accuracy could be obtained in this case through the use of 78 (*i.e.*, 3 × 26) cards.

sets of Termatrex cards are not distributed to the libraries, because it is totally impractical to attempt to keep several sets of cards up-to-date. The ability of any library to make a rapid response is provided by a priority telephone link between each library and the Corporate Technical Information Center. Several copies of the abstracts have been previously distributed and are on file in each library. Each abstract is assigned a serial number identical with the accession number of each document. The product of a search made with the optical coincidence cards is a list of the accession numbers of pertinent documents. This list is communicated by telephone to the library originating the inquiry. That library, in turn, hands the engineer a copy of each pertinent abstract, from which he selects the document he wishes to review in the original form. Problems of the acquisition of classified documents, based upon a need-to-know and security clearance, rest with the "needing" plant.

COST PER SEARCH

The average retrieval time for abstracts in response to an inquiry is 3 min., with a maximum time of 10 min., and the documents retrieved reflect better than 95% relevancy. Recently, 300 routine searches were made in a period of 1½ days; pertinent abstracts were retrieved from the file during the same period. The cost per search, including retrieval of the abstract, was approximately 11 cents. This cost does not include an allowance for the salary of the abstracters nor related overhead expenses.

The Use of Subordinate and Coordinate Indexes *vs.* the Scanning of Their Outputs*

FRED R. WHALEY
Information Systems Research Group, Battelle Memorial Institute, Washington, D. C.

Received July 6, 1964

INTRODUCTION

Users of information services, as well as the staff supplying such services, frequently scan lists of titles or abstracts or finger through cards in a card file. They may be studying lists or card files organized on a subject basis, such as a library card catalog, or lists or card files not so organized, such as those supplied by a current-awareness service. If we omit that part of their activity which is random browsing, we are left with activity that has something to do with an index. Part of such activity is actually using an index and part is studying the output of an index to decide which, if any, original documents should be procured.

It is submitted here that a better understanding of this over-all activity is possible if a distinction is made between index using and output scanning, so that each is seen in its relation to the other and to the interaction between them. This will result in system planning which is closer to system needs.

The following discussion is limited to nonnumeric systems or document retrieval, as opposed to data retrieval.

Definitions and Ground Rules.—Indexes operate in an environment consisting of:

(A) A population of items (for our purposes the file of documents or articles under consideration).

(B) A real or implied charter[1] indicating the purpose of the index and the point of view of the clientele to be served.

(C) An inquiry or interest profile, consistent with B, submitted by a client.

(D) A subpopulation of items, being a part of A, which contains information on the inquiry (C). This subpopulation may be zero even with a reasonable question.

When the index operates in the above environment, a new environmental element is created:

(E) A subpopulation of items, being a part of A, and being the output of the index when triggered by inquiry (C). Within the constraints imposed by cost, human error, linguistics, etc., E represents the index's best approximation to D.

A working definition of an index is now possible.

Within the environment as outlined above, an index is a device, which when triggered by a reasonable inquiry, points to a subpopulation of items (a small proportion of the total population of items in the collection) and indicates that most of the items in the collection dealing with this inquiry will be found in this subpopulation.

The subsequent step of examining this subpopulation for relevancy, whether done by the inquirer or by the staff of the information service or both, is held to be a post-index operation and not a part of index using.

Index Output, the Display Continuum.—Only in a hypothetical, ideal index is the output (E) for a given question identical with the actual answers (D). Real-life indexes are usually flexible enough so that inquiries can be negotiated into logical form so as either to minimize lost material with an increase in irrelevant material, as for a researcher with a "leave-no-stone-unturned" approach, or to minimize irrelevant material with some loss in wanted material, as for a practitioner wanting a good answer and quickly.

* Presented before the Division of Chemical Literature, 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1964.