can be put into the aperture card "window" is limited (we did not consider multiple cards for one item satisfactory), and the cost of the equipment necessary to use this system was prohibitive. Equipment of this type does exist in our computer section, but it is in continuous use and sufficient time could not be scheduled for our needs.

We were interested in the more sophisticated equipment using magnetic tape. Several installations using this system were visited, including the MEDLARS complex at the National Library of Medicine. However, we found that evan a small installation of the magnetic tape equipment was prohibitively costly. Neither is it practical to put entire articles and booklets on magnetic tape.

We heard about Miracode at a seminar in Florida. After their representatives demonstrated the equipment, using material from our literature collection, we concluded that Miracode could solve our problems. The cost was within our budget. The system offered rapid retrieval; it would save a great amount of storage space; it had printout capability; and it offered an opportunity to scan the collection without printing it out. Material ranging from bibliographic references and abstracts to entire books can very easily be put into the system for retrieval.

The data must be selected and coded. Herein lies the success or failure of any automated retrieval system. For this portion of the operation we use professional personnel trained in the areas of interest, primarily chemical and agronomic research as it relates to the fertilizer industry.

Miracode operates in two parts, input and retrieval. Although the input portion is necessary, its purchase is optional, as this work can be handled on a contract basis with the manufacturer, Recordak Company, a subsidiary of Eastman Kodak. The input consists of a specialized 16-mm. microfilm camera and electronic coding device, which film and code a document in one operation easily handled by clerical personnel.

The most interesting part of the Miracode equipment is the retrieval station or keyboard console. This retrieval station allows one to search a maximum of four million different subject areas or combinations of areas. The number of possible combinations depends on the number of keyboards in the console. This means that a searcher may retrieve all entries in a given subject matter area, no matter what the source, so long as they are in the Miracode system.

The Miracode equipment has even greater possibilities than we at first realized. One of our latest innovations has been the coding, filing, and retrieval of black and white and color photographs, color slides, and other visual material. The simplicity of input has made it possible for us to keep on file material of undetermined value that might otherwise have been discarded. Miracode searches at the rate of about 8000 pages per minute, or about 3000 code characters per second. Simplicity of operation eliminates the need for highly trained personnel. In fact, we hope that in the near future local patrons will carry out their own searches with minimum library assistance. Miracode also allows us to search our entire collection without regard to the divisions or categories in which it is kept. Miracode has also made it possible to acquire collections from other organizations and use them as our own. This would have been much more difficult to accomplish with hard copy or even with the more sophisticated hardware.

A Decade's Experience With A Primitive Machine Retrieval System*

FRANK S. WAGNER, Jr.
Celanese Chemical Company, Corpus Christi, Texas

Received May 31, 1966

A machine retrieval system is described which uses a superimposed coding adapted to mechanical sorting. The system has been in operation for the past 10 years. Several problems concerned with formula-functionality indexing, number of terms, etc. are examined. Most of the shortcomings are well-recognized deficiencies of all coordinate indexes.

Mechanized retrieval systems usually imply large numbers of documents or relatively complicated coding systems. Our system has neither. It is primitive in the sense that the minimal necessities are supplied by as simple means as possible. Machine methods are used to speed up and cut the cost of input; machine methods are used to recover the references needed from the file. In our decade of experience, this comparatively unsophisticated system has encountered surprisingly few difficulties. This may be in part attributed to its limited subject field, the small size of the file, or the primitive adaptability of the system; but in any case, it is perhaps instructive to examine some of the troubles we discovered and how they were managed or overcome.

* Presented before the Division of Chemical Literature, Symposium on Problems of Small Information Groups, 151st National Meeting of the American Chemical Society, Pittsburgh, Pa., March 25, 1966.

Need for some kind of mechanized information or reference retrieval system was recognized over 10 years ago. The precipitating incident was the discovery by one of our executives that our edge-notched, master index card file with over 10,000 cards was very inconvenient to sort.

After consideration of traditional subject headings, manual coordinate indexes of various kinds, and some highly sopisticated, computer-based systems, a mechanically sorted index was agreed upon. There was some interest in the possible use of document abstracts. Ultimately, however, we came to consider that the orginal document was the only abstract that would be of very much value. The high cost of abstracting entered most prominently in this decision.

The documents normally contain five to 20 pages of typescript, single-spaced with double-spacing between paragraphs. About a third of the pages consist of tables, charts, diagrams, or bibliographies, having little effect on the indexing. Ordinarily, document analysis indicates that five to 35 coordinate indexing terms be applied to each individual document.

When the system was first set up, we imagined that we would use a four-punch code for each of the descriptors and that the indexer would indicate the proper punch positions with a mark-sensed card. The regular size of a mark-sensed card allowed the use of a 10 × 27 punch field. It turned out that the indexer could not mark the card with sufficient accuracy. From 15 to 20% of the marks made on the first batch of cards were wrong, an intolerable level of error.

We cast about for ways of avoiding these errors, and found a suitable mechanical way to pick out the correct code for a particular term. The term is keypunched in alphabetic form on a card, then by using this card to search through a code deck to find the proper four-punch code, the selection is made by a relatively unerring machine process. Such a primitive device served to make the system a practical reality. We have preserved the 10 × 27 punch field, a relic of our attempt to use the mark-sensed cards for coding, in order to avoid having to redesign the remainder of the document card.

The size of the term dictionary was a source of some anxiety at the outset. As there would be too many terms in the dictionary to be coded readily in the 10 × 27 punch code field using only one punch for each term (dedicated space coding), we decided to use four punch positions to represent each term in the descriptor dictionary, thus permitting a maximum of 216,546,340 entries in our dictionary.

Statistical analyses were made of the way terms were used in other coordinate indexes at our disposal, and it was found that the names of chemical compounds formed a dominating majority of the descriptors. Furthermore, most of these chemical names were put among the descriptors to refer to only one document. Seldom did they refer to more than three documents. We had no idea of how many chemical names might be introduced eventually into our term dictionary, so we decided reluctantly to employ the empirical formula and functional fragments were quite widespread and fashionable in the mid-1950's, when we were just begining.

Names of compounds were permitted in the dictionary from the first, but they were originally precoordinated terms depicting the formula and the functionality of the compound. Misgivings about this idea began only after several years of no apparent difficulty with it. Quite naively, we tried to make a series of searches which involved distinguishing methyl propionate from ethyl acetate. Since both these compounds have the same formula and both are saturated esters, they could not be separated until some new terms were added. Later, another example cropped up. Allyl acetate, vinyl propionate, and ethyl acrylate came out as false drops in a search because each of them had the same empirical formula and possessed the same type of chemical functionality. Though the problem did not seem to be particularly pressing—at that time it seemed very simple to weed out the false drops arising from this kind of coincidence because we happen to deal with comparatively few compounds subject to this hazard—an attempt was made to clear up the names of chemical compounds in the dictionary.

A single code was introduced for each chemical name to replace the precoordinated terms. It was quickly learned that some compounds do not lend themselves readily to names suitable for listing; the systematic names are not widely known or recognized easily, and trivial ones often give a poor indication of the chemical structure. These are problems, of course, common to any coordinate index and the only solution envisioned at this juncture consists in using a simple topological notation based strictly on the structure of the compound and its constituent elements. I have developed such a topological notation in which the structure is represented by a mathematical matrix that may be capable of providing unique and unequivocal identification of even the most complex of structures. Experiments are being conducted with this notational scheme and will be described in a subsequent paper.

Another problem is related to the number of terms applied to a single document. We assumed that 35 terms were the upper limit to be applied to any single document, and 10 years have not yet produced a contradiction to this inference.

Recently, it was found that if our codes were truly random there is one chance in 100 that a duplicate code would recur when only eight terms are applied to a document. Some of the randomness of our codes is lost because we select only those four-punch codes which have four different columns represented. It has been found easier to search four different columns than it is to find more than one punch in the same column. This diminishes the number of possible codes somewhat, but not much. No problems have arisen in practice from this nonrandomness.

A number of people have experimented with the use of role indicators and links between individual terms in coordinate indexes, and we have given the matter some thought too. Though links and roles are often coupled in the literature of documentation they represent very different modes of working with problems in coordinate indexing and should be considered separately. Since our system involves superimposed coding on a common field, links between terms are practically impossible. Role indicators, on the other hand, can be applied by simply increasing the size of the term dictionary. For example

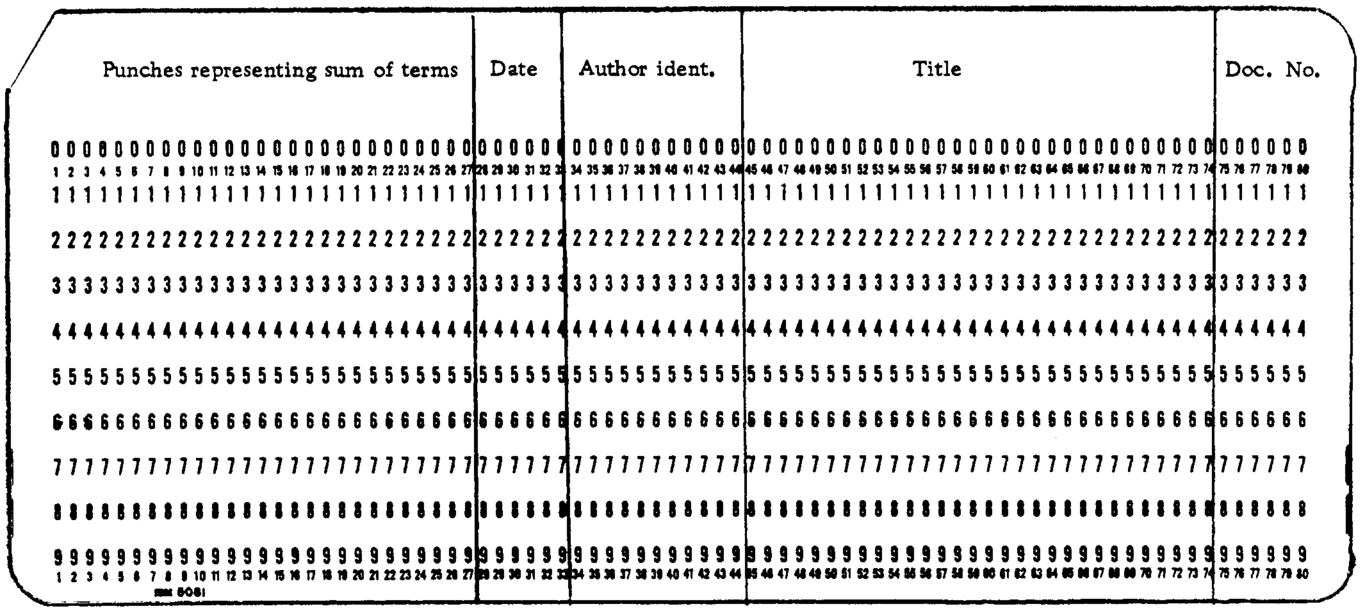

Figure 1. Search card.

the term "fractionation" may mean crystallization, distillation, or partition. Each of these can be identified by a unique code, so that separate terms for them are used. The question of whether to precoordinate terms or to give them unique codes is very difficult and is mentioned only superficially here.

As a general guideline, terms with an associated role indicator should be coded uniquely when the parent term is a peer set independent of the role indicator set. This is regrettably far easier to say than do. Since there is no risk in using up even a major fraction of the potential number of terms possible in the dictionary, the best decision is usually to avoid precoordinated terms. This tends to make the dictionary of terms more bulky, and it slows indexing, but one can overcome these difficulties partially by formulating a good thesaurus. With more and more detailed role indicators, one approaches the style of entries characteristic of traditional subject headings, and departs from the idea of concept coordination. Nevertheless, we prefer a comparatively primitive system that works rather than adhering to either the orthodoxy of coordinate indexing or the traditional subject headings.

The chief advantages of this system are largely resident in its input speed. The input is accomplished as follows: An index card is typed up which contains the document title, author identification, type of report, date, and document number. The reverse side has the terms applicable to the document, selected from the term dictionary.

These cards are prepared from the documents at a rate of 10 to 12 documents every hour by a trained literature chemist who is familiar with our dictionary. An average of about 14 terms are applied to most documents.

Each term on the back of the index card is keypunched into a description card using the alphanumeric IBM code. The information on the front of the index card is then entered into the search card (Figure 1).

The four-punched codes equivalent to the terms on the index card are then added to the description card, and all the punches are consolidated into the search card.

Retrieval is accomplished by selecting the terms to be searched, finding their codes, then sorting the entire search deck for the consolidated punch pattern. This method is somewhat better than searching the whole deck for one term (four punches) at a time. Very seldom are more than three terms coordinated for a search; usually three are sufficient for a highly selective search. Search and printout normally take 15 to 20 minutes using the IBM 082 card sorter and IBM 407 accounting machine for the printout.

This laboratory has employed a simple coordinate index with some success over a period of 10 years. Its deficiencies are recognized as largely the deficiencies of all coordinate indexes. Eventually we hope to have access to a binary computer which will eliminate the need for superimposed coding and relieve some of our anxiety about the theoretical basis of our current system. We also anticipate that a computer-based system will speed up output.

We have found this machine indexing system well suited to our current needs and expect to expand its use, with comparatively few modifications, to several different kinds of files not now being handled in a mechanized system.