A Multi-Level Retrieval System. I. A Simple Optical Coincidence Card System*

LEE N. STARKER
Warner-Lambert Research Institute, Morris Plains, N. J.

J. A. CORDERO
Lederle Laboratories Division, American Cyanamid Co., Pearl River, N.Y. 10965

Received March 11, 1968

An optical coincidence retrieval system is described that is characterized by its simplicity of maintenance and use. The system is based on the readily available IBM card. In its simplified form, it can be operated wholly by a single individual, without recourse to expensive equipment. In its more advanced form it presents certain capabilities for machine manipulation that greatly add to its flexibility.

A major problem faced by many information groups is how best to handle the retrieval problems of the individual who maintains a highly specialized file of data. In most cases, this information is of direct value to no more than a very small number of persons, and the average central information group cannot justify spending its time on the indexing, input, and retrieval of these data. It must restrict its major efforts to those activities that will serve a larger proportion of its clientele. At the same time it must be recognized that these "small" systems do have a tendency to grow. Ideally it should be possible to convert a "small" system to a "medium" system to a "large" system without having to rework any of the original input.

An associated problem arises when the information center controls a body of data that covers a wide number of users, not all of whom require access to this information at the same level of depth and sophistication. Because such systems must be designed to answer the most difficult questions, the storage form of the data tends to be complex, and one that is not always readily accessible to, or manipulative by, the casual user. This then tends either to discourage the low-level requestor, or to flood the information center with a larger number of questions, many of which could have been answered directly had a relatively simple search tool been available for use by the requestor.

We have had to face both of these problems, and have developed a general approach that has thus far proved very satisfactory for the handling of information needs for very small to very large data collections and to convert the data from one system to another. These systems are operable on an individual basis or via an information center.

THE VISUAL COLLATION CARD

The "peek-a-boo" or "optical coincidence" approach to information retrieval presents a simple, self-contained procedure, whereby an individual may develop, maintain and search an index without need to rely on a central information facility. This technique becomes even more important when there is no central information group, and any services in this direction must, of necessity, be on a "do-it-yourself" basis.

Optical coincidence searching, as we know it today, is an outgrowth of the early work of Batten (*1*), and in recent years has seen the development of the commercially available Termatrex (Jonker Corp., Gaithersburg, Md.) and Keydex (Royal-McBee Corp., New York, N. Y.) equipment that permits the maintenance of fairly large-scale systems. Another example of a large-scale system is the "Microcite" (*2*) development at the National Bureau of Standards.

The purpose of this paper, however, is to describe the development of a simple card format that permits the use of optical coincidence searching at a very elementary level, and without the use of expensive drilling apparatus.

We had at one time considered the IBM Port-a-Punch (International Business Machines Corp., New York, N. Y.) card for this use, but found that there were sufficient inconveniences in this application to warrant an improved development.

Since the IBM card was inexpensive and readily available, we decided to attempt the elimination of some of the problems inherent in Port-a-Punch cards, but still retain the IBM card format.

We found particularly annoying the fact that a Port-a-Punch deck would only serve for 480 documents, and secondly that punching out the pre-scored holes was not always as clean and simple an operation as might be expected. The punch-out process often resulted in ragged holes, or left pieces of the card dangling from the perforation. These problems were further aggravated when attempts were made to reproduce decks of these cards. The irregularity of the holes often resulted in non-reproduced punches, as well as jams and torn cards.

Our first answer to these problems was the design of a "Visual Collation" card that permitted the control of 960 documents per deck of term cards. This was achieved (Figure 1) by assigning all punch positions on the card so that each one could be used to represent a document number, thereby doubling the number of records that could be handled with a Port-a-Punch system.

In this approach, documents 1 through 9 are recorded by a 12-punch in card columns (cc) 1 through 9, while documents 10 through 809 are punched in the body of

*Presented in part before the Third Middle Atlantic Regional Meeting, ACS, Philadelphia, Pa., February 1968.

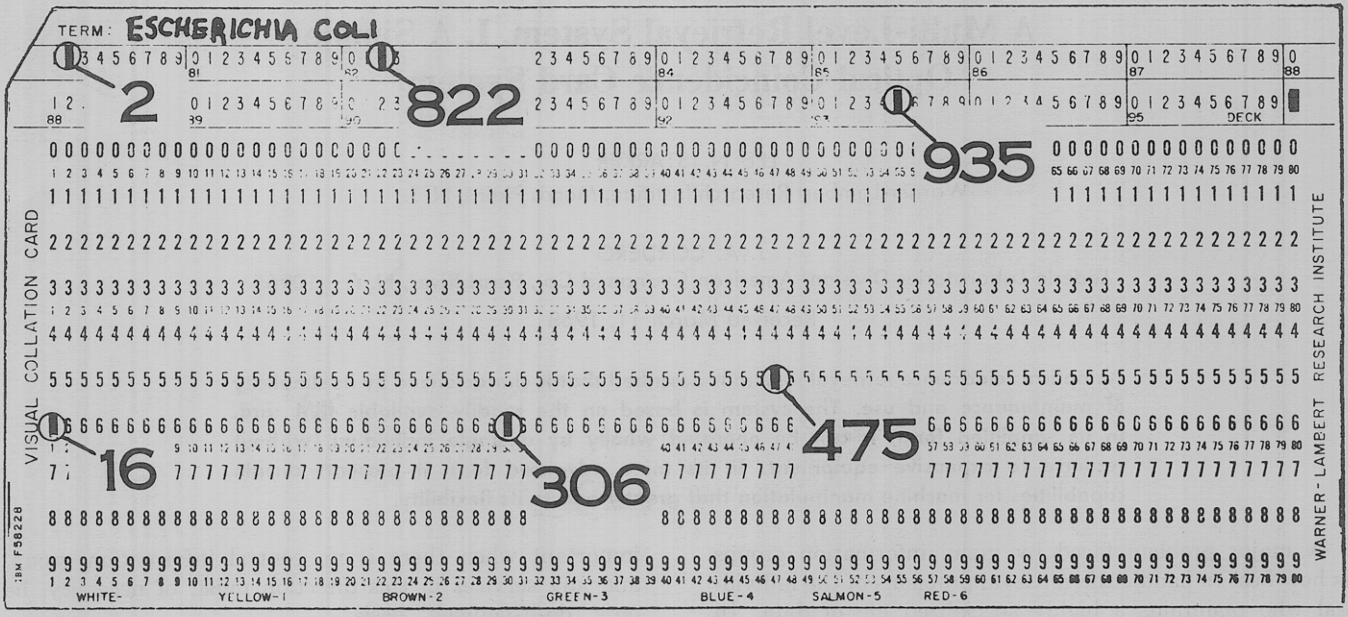


Figure 1. Visual collation card

the card. Use of a simple mnemonic numbering device, in which the *last* digit of each document number represents the *row* of the IBM card to be punched, while the first, or first two digits, represent the column to be punched, permits easy read-out from these cards. Thus document 12 is recorded in cc 1, row 2; document 67 is punched in cc 6, row 7, etc. (see Table I for additional examples). Document numbers above 809 were provided for in the 11- and 12-punch positions over cc 10 through 80. The punch positions for these documents are not completely mnemonic, and it is easiest to simply follow the printed arrangement on the card.

We now have a deck of cards that will handle almost a thousand documents. We did not think that it would be particularly disconcerting to eliminate the use of document numbers between 961 and 999 when a large file was to be handled. Thus we number from 1 through 960 for deck 1 (skip 961 through 1000), and then number from 1001 through 1960 for deck 2, etc. In such instances, deck 2 was prepared on yellow cards, which was an immediate signal that all document numbers in that deck required that 1000 be added to them. Thus the document punch in cc 6, row 5 of the white deck is read as document 65, while the same punch in the yellow deck represents document number 1065. If a third deck was to be used, it would be placed on brown cards, and the punch in cc 6, row 5 would then represent document number 2065.

Table I. Document Number Punch Codes

Document Number	Visual Collation Card		Visual Collation Variable Term-Field Card	
	Column	Row	Column	Row
2	2	12	2	11
4	4	12	4	11
8	8	12	8	11
16	1	6	1	6
23	2	3	2	3
85	8	5	8	5
306	30	6	30	6
475	47	5	47	5
710	71	0	...	...
822	22	12	...	...
916	36	11	...	...
935	55	11	...	...

This card now made it possible to set up a peek-a-boo system, with the assignment in the usual way, of one index term per card. These terms are either written or typed in the upper left-hand portion of the card. Input to the deck could be carried out with a standard IBM keypunch if desired or, as was more usual, by the individual directly concerned, using a conductor's hand-punch (Bonney-Vehslage Tool Co., Newark, N. J., No. 19½ punch, 4-inch throat, die No. 29) fitted with the rectangular IBM type of die (Figure 2). Very little care is required to align the punch properly so that the resultant hole is centered—and as many as four to five cards can be punched easily in one operation.

Searching, of course, is carried out by selecting the appropriate term cards, superimposing them, and reading out those document numbers through which light will pass—i.e., numbers that are common to all of the selected

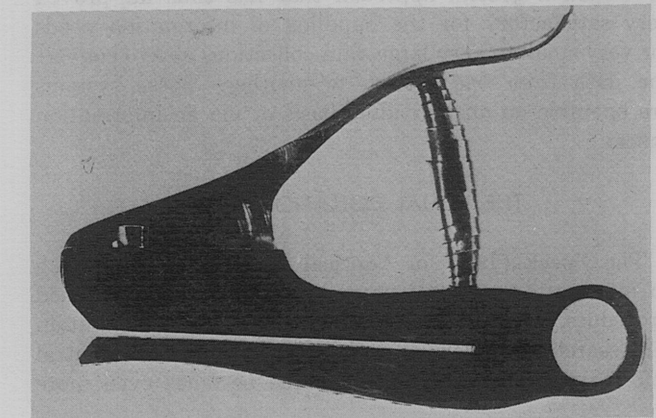


Figure 2. Hand punch

cards. Because of the relatively large size of the hole, we did not find that a light box was necessary. The relevant document numbers could be determined easily by superimposing the cards while sitting at one's desk. If desired, they could be held up to the light in the room to ease the identification procedure.

Interestingly enough, a permanent record can be made of the search output by placing the selected superimposed term cards on a Xerox copier and preparing a copy of the result. The common holes show up as dark areas on the card image, and can be easily read.

We did find the process of superimposing the cards was greatly facilitated by pre-punching cc 80/11 in all cards. This always assured us of having a punch common to all cards, and made it easier to line up the term cards involved in any search question.

This "line-up" punch meant that we could now control only 959 documents per deck rather than 960, but this did not appear to be too high a price to pay for the added convenience (Figure 3).

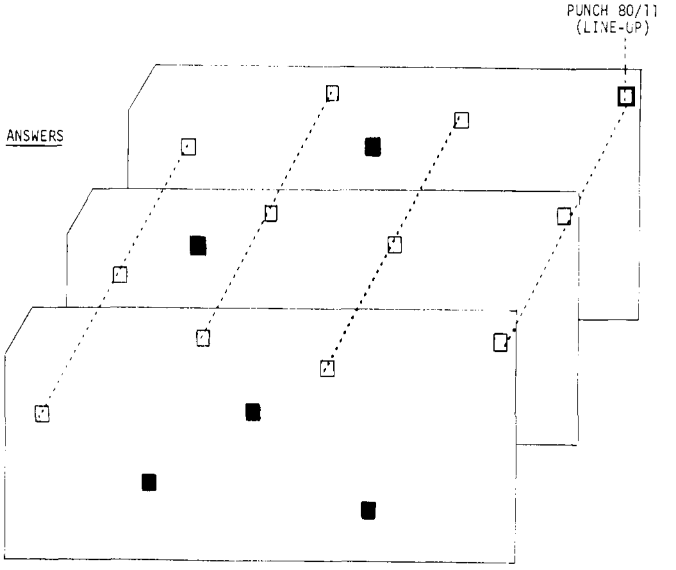


Figure 3. 80/11 "line-up" punch

The Visual Collation cards worked well in most respects, but annoyances were encountered as the cards began to wear, or when applications developed where it was desirable to prepare multiple copies of a given deck.

The actual reproduction of the punched master deck was easily carried out with machine-punched cards, and usually was not a problem with hand-punched cards, if minimum care had been exercised in the initial preparation of the master deck. Problems did arise, however, when it came to copying the term names on each card of the new decks. This required a manual procedure, with extreme care being necessary to record always the proper term name on the corresponding duplicate card.

To remove this degree of inefficiency, and to add several additional degrees of flexibility to the system, the 960-document Visual Collation card was then redesigned to provide a vehicle that could be completely machine processible.

THE VISUAL COLLATION VARIABLE TERM-FIELD CARD

The resultant product has been termed a "Visual Collation Variable Term-Field" card (Figure 4). Although this card can still be used as is the Visual Collation card, it is now limited to a total of 809 documents per deck, since no provision is made in the 11-, 12-punch area for document numbers 810 through 960. Document numbers 1 through 9 are now recorded by using the 11-punch in cc 1 through 9, while the "line-up" punch has been placed in cc 80, row 12. The major differences lie in the way in which this card is used.

To obtain the maximum effect from this variation, the recording of document numbers is restricted to cc 1 through 50. By further restricting the use of cc 50 to the 0-punch only, these new cards will now control 500 documents per deck. Thus cc 51 through 79 are available

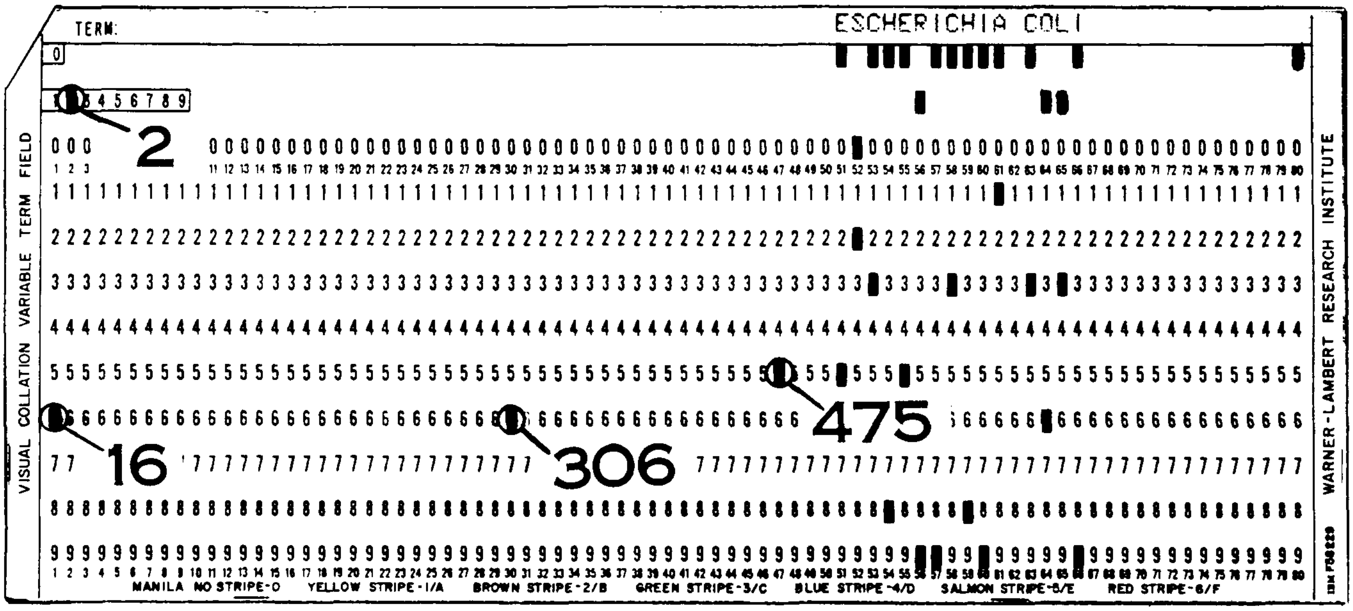


Figure 4. Visual collation variable term–field card

for the index term, which can now be keypunched in this field for machine handling.

Obviously, this kind of input now requires that the user have some access to a keypunch. However, once, the terms have been punched into the cards, the deck can still be handled as was the 960-document card described earlier.

This new card now has the advantage of being completely machine-reproducible with respect to both the document numbers in cc 1 through 50 and the term name in cc 51 through 79, thereby removing one of the major disadvantages described for the earlier card. Interpretation of the term name field of the reproduced deck, of course, is required to make the name readily legible. In addition, the document numbers are now completely mnemonic, which greatly assists both input and read-out.

The device of limiting the punching of document numbers to no farther than cc 50, row 0 (document 500) simplifies the use of multiple decks for collections of over 500 documents (and removes the necessity for omitting any numbers in the document sequence). The first deck now covers documents 1 through 500, the second deck controls documents 501 through 1000, the third deck handles documents 1001 through 1500, etc. With this system, a punch in cc 9, row 3 identifies document number 93 in deck 1, document 593 in deck 2, document 1093 in deck 3, etc. In addition, each deck can be differentiated, not only by the use of a differently colored card or color stripe, but by machine as well. This is accomplished by entering an identifying punch in cc 80. Thus a punch in cc 80/1 identifies deck 1 (1 through 500), 80/2 identifies deck 2 (501 through 1000), etc. Reproduction of the cards will now carry through all necessary information—the document numbers, the term name, and the deck number—to determine completely the portion of the file to which it belongs.

Where the user has access to any kind of printout equipment, he can make use of an additional tool that will greatly assist his input of new document numbers.

The recommended procedure is to alphabetize the term cards—either by machine or manually—and then prepare a listing from the term field (cc 51 through 79) of these cards. This printout (Figure 5) can be used as an indexing worksheet by entering opposite each term the document numbers of any records that are found to be applicable. Once a group of documents have been read and indexed, the document numbers can be entered on the appropriate term card. Since the cards and the index terms are in the same alphabetical sequence, each term card need be handled only once for a given group of documents, and can be replaced in the file immediately without the time-consuming find and refile procedures that would be necessary if all terms for each document were to be punched at one time.

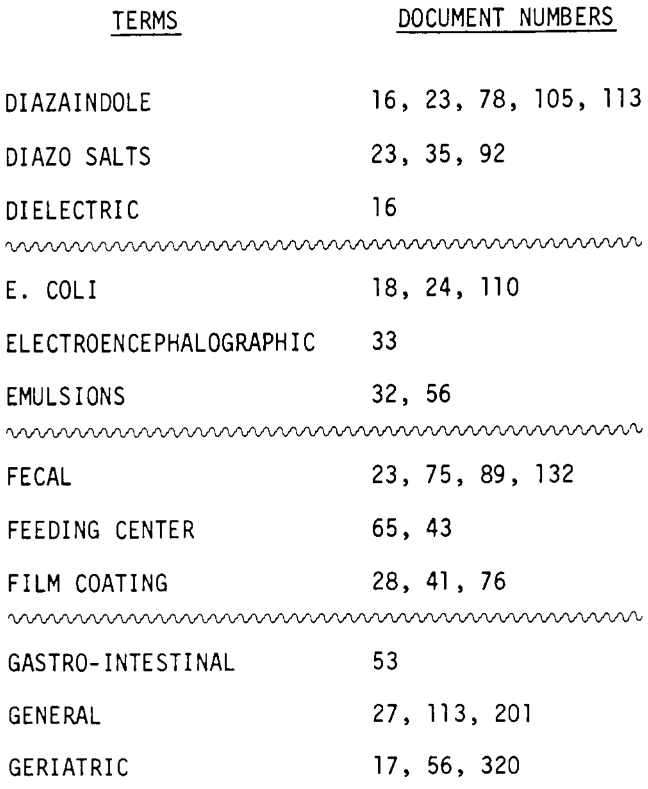
These Visual Collation Variable Term-Field cards have been well received, and over the past eight years have been used in a wide number of applications. Since most of the applications have been built on personal collections of reprints or data, the individual concerned has developed and maintained his own dictionary of terms, determined his own indexing policies, and carried out his own searches. The operational flow is shown in Figure 6.

Information Center personnel have been involved only to the extent of providing consultative assistance and arranging for keypunching (where desired), card reproduction, and term printouts.

Of particular interest has been the development of several systems where a single individual organized a retrieval approach for a collection of reprints which were of some interest to other staff members. In such instances, reproduced decks of the term cards served to make entry to the system readily available to all individuals concerned. As the master deck was periodically updated, it was reproduced and distributed. On receipt of each new deck of cards, the outdated decks were simply discarded.

Document collections in the areas of biochemistry, cosmetics and toiletries, and sulfonamides, as well as library bibliographies, have been controlled with this retrieval tool. One of the earliest applications was to record the properties reported in the literature for new antibiotics so that as additional materials were isolated and purified, ready comparisons could be made, and the novelty of the new isolate could be established. Term vocabularies have ranged typically from 50 to about 200 words. These systems tend to be user-controlled and, therefore, the depth of the indexing need not be too great, since the searcher will already be relatively familiar with the body of literature that is under consideration.

The number of decks that can be easily searched is almost inversely proportional to the frequency of searches that will be carried out. This follows because every group of 500 papers requires a new set of cards, and every complete search must then be carried out through each deck of cards. A system controlling 4000 references (the largest that we have seen) would call for eight sets of



TERMS	DOCUMENT NUMBERS
DIAZAINDOLE	16, 23, 78, 105, 113
DIAZO SALTS	23, 35, 92
DIELECTRIC	16
E. COLI	18, 24, 110
ELECTROENCEPHALOGRAPHIC	33
EMULSIONS	32, 56
FECAL	23, 75, 89, 132
FEEDING CENTER	65, 43
FILM COATING	28, 41, 76
GASTRO-INTESTINAL	53
GENERAL	27, 113, 201
GERIATRIC	17, 56, 320

Figure 5. Index work sheet

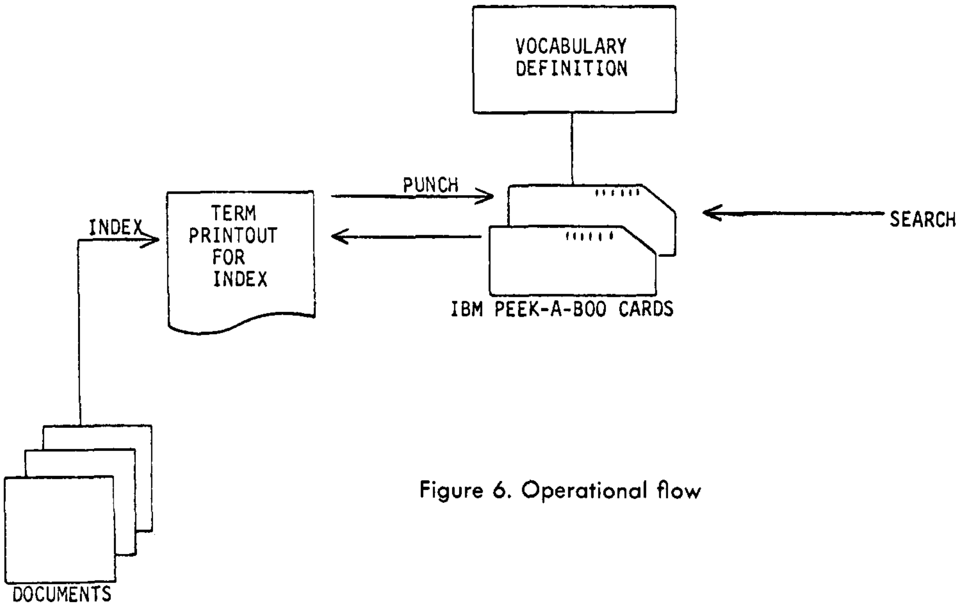


Figure 6. Operational flow

Visual Collation Variable Term–Field cards, and therefore every search would have to be carried through all eight decks.

We have not attempted to establish any definitive guidelines as to when a system of this type can be expected to become too cumbersome for efficient handling. A good rule of thumb seems to be that when the number of daily searches times the number of decks exceeds ten, thought should be given to a new, more effective search tool.

Paper II of this series will describe the techniques that we have developed for converting retrieval systems based on the Variable Term–Field cards to systems that are more appropriate for larger collections and/or more frequent searching.

LITERATURE CITED

(1) Batten, W. E., "Punched Cards," R. S. Casey and J. W. Perry, Eds., pp. 169–181, Reinhold Publishing Co., New York, N. Y., 1951.

(2) Wildhack, W. A., and Stern, J., "Punched Cards," R. S. Casey, J. W. Perry, M. M. Berry, and A Kent, Eds. pp. 147–150, Reinhold Publishing Co., New York, N. Y., 1958.

Computer Indexing of Polymer Patents*

M. M. DUFFEY, I. M. KLANBERG, S. C. MAHR, L. L. MEIER, and J. L. ROMSTAD
E. I. du Pont de Nemours and Co., Experimental Station, Wilmington, Del. 19898

Received March 7, 1968

A computerized system is described for indexing polymers using link-role numbers to describe the polymer backbone (structure), to distinguish between homopolymers and copolymers, to distinguish between unmodified and modified polymers; and to differentiate in addition copolymers between comonomers that must be present and a set of alternate comonomers—*e.g.*, in the case where *x* is copolymerized with *y* or *z*. In all categories the roles also indicate whether the polymer is a product, a reactant, or present. Generic polymer terms are added to cut costs in generic searches.

The Central Research Department in Du Pont has a patent library which houses 600,000 to 700,000 indexed patents—both U.S. and foreign. Included is a comprehensive collection of British, French, German, and U.S. patents which deal with the preparation and use of polymers.

In the early thirties our polymer indexing system depended on buying extra copies of patents and filing these under the name of each monomer of an addition polymer or under the name of each repeating group of a condensation polymer. In 1952 we started making abstract cards for each patent and filing these cards in alphabetical order by headings listed on them. The number of abstract cards per patent varied depending on the number of headings needed to describe the subject material in the patent. Abstract cards relating to addition polymerization had headings listing each monomer in the polymer. The order of monomers was changed on subsequent

*Presented in Symposium on "Automation of Information Operations," ACS Middle Atlantic Regional Meeting, Philadelphia, Pa., February 1, 1968.