Economic Evaluation.—What about the cost of providing a service such as we have discussed? There appear to be four major factors which interact to effect the time and cost of a search. They are: (1) the total number of terms regardless of how the terms are distributed within the requests; (2) the number of references in the file, and the number of index entries which provide access to the file; (3) the number of references in the file which satisfy the terms of the search; and (4) familiarity of the searcher with the vocabulary of the search.

Our experiments included a number of human searches, either parallel to or similar to the machine searches we conducted. Although we do not yet have enough data to draw final conclusions, the most favorable combinations of the four parameters do not appear to justify the present cost of machine search. However, considerations of speed, available staff, and the alternative uses for the time consumed if each scientist were required to make his own searches may justify the cost for machine searching. The cost of operating second-generation systems we contemplate may be less than that of a manual system.

Future Development.—A number of possible improvements which would facilitate selection of relevant titles have been suggested by participants in these experiments. These are detailed in another paper.[5]

CONCLUSION

In summary, we have experimented with automatic retrieval of references from *Chemical Titles*. The results have provided a current-literature alerting system at Eli Lilly and Co., a current awareness system and a retrospective search file at Olin Mathieson Chemical Corporation, and indexed annual bibliographies for various areas of chemistry at the Chemical Abstracts Service. Comparisons with human searches of *Chemical Titles* suggest that the present experimental machine systems are more expensive to use under certain conditions. Advantages gained from speed and utility of the search output, and prospects for improved programs, indicate that systems which are more practical and economical than those now in use will evolve.

ACKNOWLEDGMENTS

The authors wish to express their particular thanks to Mrs. J. B. Haglind of Olin Mathieson Chemical Corporation; and to Mrs. Claudene Frank, Mr. Karl Hardey, and Mr. William Muirhead of Eli Lilly and Co. for their contributions to the experiments described herein.

REFERENCES

(1) See R. E. Maizell, *Rev. Doc.*, **27**, 106 (1960).
(2) R. R. Freeman, and G. M. Dyson, *J. Chem. Doc.*, **3**, 16 (1963).
(3) Unpublished readership survey, July, 1962.
(4) H. P. Luhn, *Am. Doc.*, **12**, 131 (1961).
(5) R. R. Freeman in "Automation and Scientific Communication," H. P. Luhn, Ed., Papers contributed to the Theme Sessions of the 26th Annual Meeting of the American Documentation Institute, Washington, D. C., 1963, part 2, pp. 213–214.

Some Unusual Features of a Chemical Retrieval System Used in the Eastman Kodak Company*

By CARL R. HAEFELE and JOHN F. TINKER
Research Laboratories, Eastman Kodak Company, Rochester, New York
Received April 22, 1963

This paper discusses briefly the mechanical and manipulative aspects of a system used in the Kodak Research Laboratories to index and retrieve chemical information. This system is used to locate the information on specific compounds and to show where samples can be obtained within the Company. Just as the system can uncover a single compound, it can also be used to retrieve all compounds of a given class, *i.e.*, with a given functional group.

Some features of our system are common to other systems, but there are certain aspects which appear to us to be unique. We claim no priority for these innovations, but we have not encountered them elsewhere, and so we present them as a new approach to the handling of chemical compounds.

The large number of organic compounds synthesized and used in the Kodak organization will inevitably require the use of high-speed computers for efficient, economical retrieval of information. However, the system has been developed so that the initial work can be done with simple sorter-collator equipment until the volume of data makes this approach impractical. When this happens, a conversion will be made to Minicard, computers, or some other type of high-speed hardware. This means that information entered in punched cards for use on the sorter-collator equipment has had to be in a format that could be accepted by a computer. This restriction created some limitations; random superimposed coding, or multiple punches, in columns that would be unintelligible to a computer were precluded. These limitations forced us to abandon the possibility of entering all chemical structural data for a single compound on one card. Once we decided

* Presented before the Division of Chemical Literature, 142nd National ACS Meeting, Atlantic City, N. J., September 12, 1962.

to use more than one card per compound, we had great freedom. We were not bound or hampered by the capacity inherent in one card; the system could be very flexible, open-ended, and capable of handling the unanticipated problems that might occur as the system developed.

A system was devised in which various aspects of chemical structure are separated into several major groupings. Each major grouping is assigned to a separate card, and specific columns in that card are assigned to the various functional groups that may be a part of the major group. For example, only oxygen functions outside of rings are recorded on one card, and specific columns on that card are reserved for specific functional groups as carbon–oxygen, oxygen–hydrogen, or carbon–oxygen–hydrogen. As it was finally worked out, 17 cards can completely describe the structural configuration of any compound. There is considerable unused area on the cards for future expansion; and if more space is required in the future, it will be a simple matter to create new cards. You can picture the array of file cabinets that would be required if 17 cards were used for each of 500,000 compounds. However, it's not as bad as it seems. The average number of cards used per compound is 7 or 8, and when we reach 500,000 compounds, the information will be in the computer, anyway.

How does the system work? The various cards describing each type of major segment are defined by letters A through R, omitting I and Q. Each letter is punched in a filing field for sorting into the respective chemical groups for easy filing. Thus all A cards for all compounds in the system are together, as are the B cards, etc. Each compound is assigned a unique accession number or Company number which identifies that compound in its travels around the Company. All test data and report descriptions use this unique number. The number is punched in a fixed field on every card required to describe the particular compound.

Figure 1 shows the information recorded on each type of card.

Card A contains the molecular formula.
Card B is for the number and type of ring structures.
Cards C, D, E, F are for the size and number of heterocyclic rings.
Card G has fixed fields for the number of carbon atoms in the longest straight chain in the molecule. It also has fixed fields for alkyl groups and unsaturated linkages.
Card H is reserved for oxygen functions formed outside of rings and has a column reserved for each one. For example, a punch in column 18 of card H indicates the presence of an OH group. In discussing these functional groups, it is understood that they are made up of carbon and/or hydrogen atoms in addition to oxygen.
Cards J through P contain other types of functional groups such as nitrogen, sulfur, etc.
Card R shows the Patterson ring number and some ion codes, such as sulfate, *p*-toluenesulfonate, etc.

The inverted system of filing is used to file the large number of cards. Two of the cards required to describe *o*-aminophenol are H and J. Card J has the amino function and card H has the phenol function. The amino in card J is punched in a column reserved for amino functions; the phenol function in card H is punched in the column reserved for this function. The J card is filed with all the other J cards and the H card is filed with all the H cards. To retrieve *o*-aminophenol, we pull all the J cards and all the H cards (not touching any of the other card decks, A, B, C, etc.). We now sort the J cards in the column reserved for amino, and the H cards in the column for phenol. Thus we have two piles of cards: one containing amino and the other containing phenol. These are matched with a collator for an accession number in common. The accession numbers that match are those of compounds containing amino and phenol. The example just described is of course an extremely simple one.

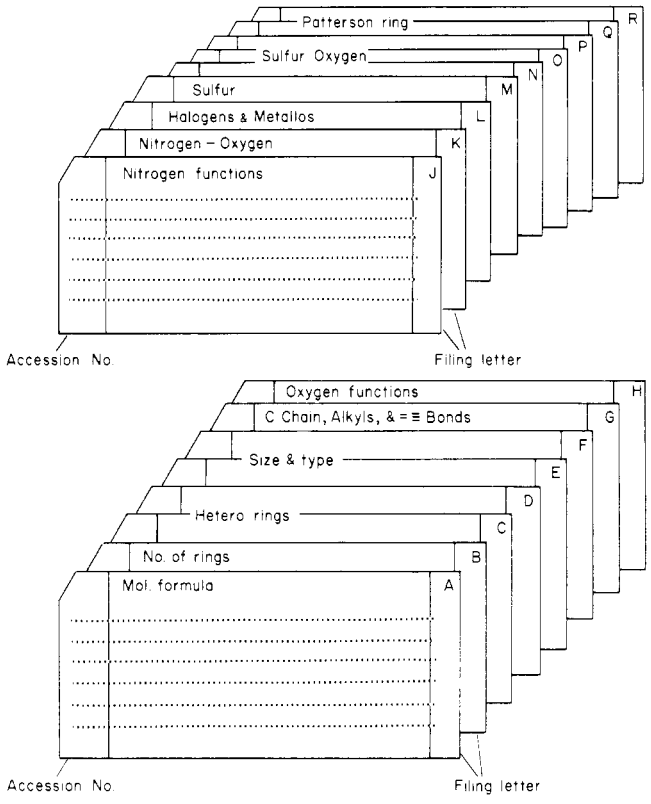

Fig. 1.—Multi-cards for functional groups.

If any group of cards, such as the J group, becomes too voluminous to handle conveniently, it can be "file-expanded" or divided into subdecks; each subdeck contains a single function, as, for example, amino. Thus, when we are looking for an amino function, this specific subdeck of the J group is removed and collated against the subdecks of other functions being retrieved.

We should like to re-emphasize the fact that an entire column in a card is reserved for one functional group. This means that a punch anywhere in the column is sufficient to indicate the presence of a given functional group in a compound. Considerable additional information concerning the functional group can be obtained by the position of the punch within the column. One of the parameters described by such a positional punch is the manner in which the functional group is connected to the other atoms in the molecule. We considered, first, showing every connection, but we found that we were getting very close to a linear notation, and this involved so many rules and regulations that we decided to show only the most important connection. Accordingly, a list of connectors was drawn in order of priority and a punch code was provided to show this connection.

Connected to anything other than N, S, C or halogen	1
Connected to nitrogen in a ring	2
Connected to sulfur in a ring	3
Connected to carbon in a ring	4
Connected to nitrogen *not* in a ring	5
Connected to sulfur *not* in a ring	6
Connected to carbon *not* in a ring	7
Connected to halogen	8
Uncertain connector	9

To illustrate how well the list and punch code work, suppose we wish to differentiate the OH group of an alcohol from the OH group of a phenol. The presence of the OH group is indicated by punches in column 18 of card H. The OH group of an alcohol is connected to a carbon not in a ring. This is shown in the connector table to a 7 punch, so a punch would be put in the 7 position of column 18. The OH of a phenol is connected to a carbon in a ring which is indicated by a 4 punch. With one pass through a sorter, aromatic alcohols can be separated from aliphatic alcohols.

The zone punches are used to show still another parameter, that is, the number of times a functional group connected in a particular manner is present in the molecule.

The presence of only one functional group is not indicated other than by its connector.

The presence of two functional groups is indicated by a zero punch.

The presence of three functional groups is indicated by an 11 (−) punch.

The presence of four or more functional groups is indicated by a 12 (+) punch.

Suppose we wish to differentiate between a hydroquinone and a diol. The hydroquinone would have a punch in column 18 of card H in the 4 position to show that the OH groups are connected to a carbon in a ring, and a zero punch in column 18 to show the presence of two OH groups. The combination of a zero and a 4 in the same column makes the letter U. To indicate the diol groups, the punches would be a 7 to show the connection to carbon not in a ring and a zero punch to show the two OH groups. This combination makes the letter X. Thus, with two punches in one column, we are able to describe three parameters: first, that the functional group is present; second, how it is connected; and third, the number of times it is connected in this manner. The information in this format is also compatible with computer usage.

This elaboration permits us to file-expand still further our original deck H cards. The deck can first be expanded into subdecks on the basis of the individual columns or functions, further divided on the basis of the connector punch, and still further on the basis of the connector multiple combination punches. Therefore, a great many compounds can be handled in an inverted file with this type of file expansion before the decklets become too cumbersome to handle with a collator system.

Up to this point, we have said that each of the various cards contains only the accession number of the compound, the functional group designation, and the file letter showing the card type. This leaves a great deal of space in the card for other data. Some fixed fields have been set aside for punches indicating the nature of the compound, such as natural polymer, synthetic polymer, salt, etc. Another field tells the class of compound: heterocyclic, aliphatic, alicyclic, or any combination thereof. Another field tells the source of its origin: chemical laboratory or plant. This information is repeated in all the cards used for each compound. Although such information is very general, it has been helpful. It can be used as a starting point from which to eliminate those cards which do not apply to the particular request.

One other helpful innovation is the use of an alphabetical cross index in each card used to describe a compound. As mentioned earlier, each different type of card is identified by a specific letter according to the kinds of information located on the various cards. The letters of all the cards used to describe a given compound are *all* punched into *all* of the cards used to index that compound. For example, let us describe *o*-aminophenol. Four cards are required to describe this compound: card A, to show the molecular formula; card B, to show a single, unfused aromatic ring; card H, to show one OH group connected to a carbon in a ring; card J, to show one primary amine connected to a carbon in a ring. The four letters A, B, H, and J are punched in all four cards in the field for the alphabetical cross-index code (Fig. 2). They are separated for filing by sorting for each of these letters in the filing field.

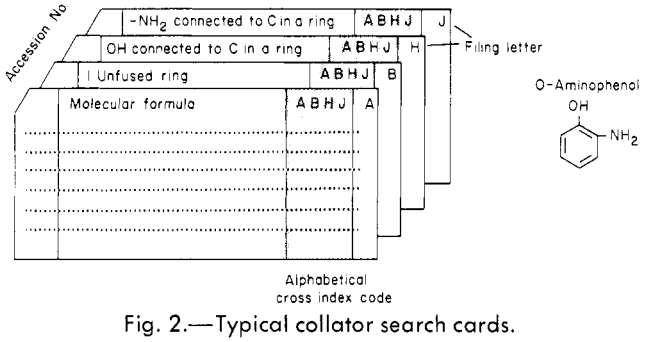

Fig. 2.—Typical collator search cards.

The alphabetical cross-index code is used as a rapid means of eliminating all cards which could not possibly contain the answer to a request. A search for simple aminophenols would uncover many complex compounds containing many other functional groups. These groups would be located in other cards, and the identifying letters of those cards would be in the alpha cross index. To avoid the retrieval of a mass of unwanted compounds, a search can be made on the alpha codes for the undesirable functional groups. For example, to eliminate any aminophenols containing S atoms, all cards containing sulfur functional groups, that is, cards identified by the letter M, N, O, or P, could be retrieved and eliminated as not being relevant to the search for simple aminophenols. The search strategy varies with each request and must be designed so that it will eliminate all cards that could not contain the desired fragments.

We shall now trace the logical sequence of steps in a typical search. We want all primary aminophenols. The required functional groups are present in cards B, H, and J. Card B contains a column showing a single, unfused aromatic ring. This decklet is removed from the file and sorted for the presence of an H and J in the alpha cross index. The discards could not possibly fulfill the request. The decklet for one OH group connected to a carbon in a ring is removed from the H-card file. It is sorted for the

presence of the letters B and J. The same procedure is used for the amine function in the card-J deck. It is sorted for the presence of B and H in the cross index.

Each specific functional group required was initially selected and searched for the possible presence of the other two required groups. These three refined decklets are then matched for common accession numbers. The results fulfill the request as all the knowns are present in the three cards and only the cards that can meet the three requirements are matched. The others have been discarded in the alpha cross-index search. It is amazing how efficient the alpha cross-index technique is in reducing the number of irrelevant cards.

Using the alphabetical cross-index code and the file-expanded inverted file of functional groups, we believe that many thousands of compounds can be placed in the system before it becomes too unwieldy for sorter-collator searching.

The main features of the system are that it is a multi-card system; it can be used with simple, punched-card equipment, yet is compatible with computers; it is open-ended and flexible; and it requires no highly trained coding specialist or expensive machinery.

The clerical details involved in operating this system are as follows.

The chemist who has synthesized a new compound is asked to fill out a 4 × 6-in. work card in the way he thinks of the compound, using his nomenclature and his version of the structural formula. He also includes the molecular formula, the name of his department, his notebook number, and any other pertinent data, such as boiling point or melting point, that he may have. The accession number which will identify that particular compound in the future is assigned. On the back of this 4 × 6-in. card is printed a form for dates and witnesses for legal protection and patent applications.

From this card a neat 3 × 5-in. card version is prepared. The 3 × 5-in. card is of great use and is duplicated many times. Card files arranged by accession number and by molecular formula have been established in many different plants and laboratories throughout the Company. These enable the chemists in widely separated areas to have access to these files. By consulting these files, a chemist can quickly obtain information on compounds that answer his search requests and learn where they may be found. Updated and new compound cards are distributed to all these files. A list of new compounds is published using the 3 × 5-in. cards and photographing eight cards to a page. Lists are issued when approximately 100 new compound cards have been accumulated. This list is sent to various chemical groups within the Company to alert them to the existence of new compounds. The cards are also reproduced on gum-label stock for pasting in notebooks, on sample bottles, etc., to eliminate the redrawing of structures and other sources of clerical errors.

The compound is entered into the index by breaking its structure into structural fragments and entering these fragments on a code sheet. This is just like taking apart a jigsaw puzzle. The intellectual effort is supplied by chemists who can easily and accurately describe the chemical fragments with ten minutes instruction.

The fragments are encoded *via* a dictionary into alpha-numeric computer codes. The accession number is punched into a fixed field on an IBM card, followed by free-field punching of the molecular formula and alphanumeric computer codes. This type of free-field punching makes for very easy and rapid keypunching. The card is then fed into a computer. The computer rearranges the various codes according to a program and produces an output card that is suitable for entering onto magnetic tape when the need arises. It also produces the many fixed-field, collator search cards that we have described as the heart of our present system. For those who do not have computers handy, the sorter-collator search cards can be keypunched directly. We have found that the lack of uniform fields and the relatively few punches per card have made punching very slow and subject to error. Using a computer costs only four cents per compound, or less than a penny per search card. Another advantage was the editing and checking for errors that were done at the same time.

Figure 3 is an over-all schematic view of the present system. From a 4 × 6-in. card, a 3 × 5-in. card is prepared and used for (1) multiple files throughout the Company, (2) lists of new compounds distributed to the chemists, and (3) the preparation of gummed labels.

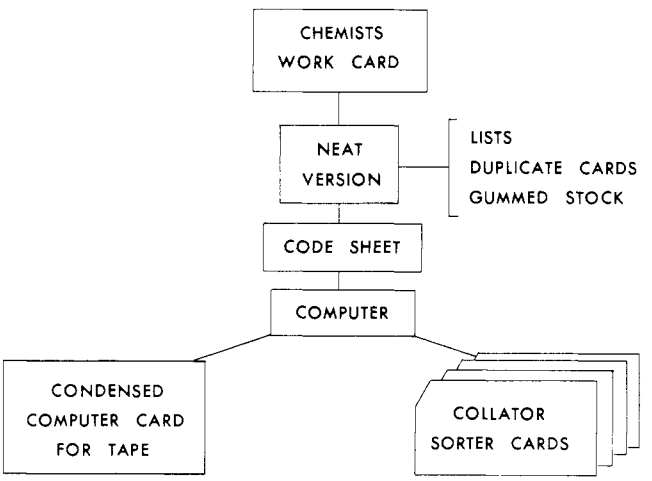

Fig. 3.—Flow chart of chemical retrieval system.

The chemicals are coded according to structure and functional groups and these are encoded and keypunched for computer use. The computer produces condensed cards for entry onto magnetic tape as well as fixed-field search cards for the sorter-collator equipment. The system is relatively new, and hence we have not had enough experience to evaluate its usefulness properly, but we are confident from results obtained to date that it will be of considerable help to our chemists.