Use of Machine Methods at Chemical Abstracts Service*

By G. M. DYSON AND ELIZABETH F. RILEY**

Received August 24, 1961

Several years ago, Chemical Abstracts Service embarked on an intensive study of mechanization. The ultimate goals are reduction of costs, speeding of indexes, and making available certain correlations not easily done from conventional indexes. To that end, an immediate goal is automation of all repetitive hand operations and all repetitive logic operations.

Among the most challenging projects are those concerned with automation of the logic operations of literature searching.

The first step was preparation of a "Super-Collective Formula Index" including all compounds ever indexed in Chemical Abstracts. Two copies of every existing formula index were marked, cut up, and pasted (still in blocks by molecular formula) on McBee Keysort cards. Since the initial cards are a work pack and are more or less expendable, the cheapest 5 × 8 sulfite pulp cards were used. The McBee cards are die-punched and printed, on all four margins, with areas of labeled number holes for the various atoms that may appear in the molecular formula. Two operators punch out the correct numbers for the molecular formulas by use of McBee 6103 Electric Punches. The champion paster can turn out 2900 cards per day; the champion puncher easily does 8000 cards per day. From January 18, 1959, to date, the girls have completed 834,000 cards, which fill 448 file drawers. The file is complete through the formula index last-issued (Volume 54*a*) and entries from Volume 54*b* are currently being cut. The slowest step is hand verifying of the McBee cards; no operable machine verifier for McBee cards is available.

The initial pack of McBee cards is needle-sorted, by subject, with a view of giving each chemist who works on the cards the subject matter along the lines of his interest. Each special deck, thus segregated, is arranged in sequence by formula. Once sequenced, all entries on the same compound fall out together. At this point the task of breaking down the block entries and making entries by individual compound is done, and summary cards are made for each individual compound. The summary cards are also McBee Keysort cards, but they are of somewhat better stock so that they will take ink lines. They are a different shade, so that visual recognition will be instantaneous.

The cipher for each compound is entered on the proper summary card, and data and information about the compound are accumulated. A register number is assigned, once and for all, and is henceforth used to signal the computer for data on that compound. Both types of McBee cards are shown in Fig. 1. Either the initial deck or the summary deck can be used as a direct access file if necessary.

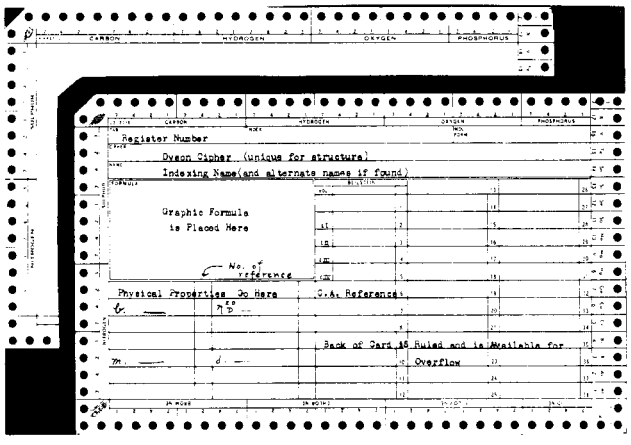

Figure 1. Two types of McBee Keysort Cards

Cards from the summary deck are used as source cards for the preparation of eigth types of decks of IBM Cards. All of the IBM cards have been designed for separate printing now, with an IBM 866 Typewriter with Document Writing Feature, and for computer use to make tapes for the IBM 1401.

Type 1 Deck. Dyson[1] has described the initial IBM deck, which bears the register number, the cipher, and the molecular formula (or code therefor of the compound. It should be explained that the sample cards (Fig. 2) are matched to the small printer, and the apparently extraneous signs are signals to the machine; they cause shifting, underlining, etc. The IBM 1401 can recognize the same signals and produce like text. A period (.) causes the next character to print as a subscript (if a number) or in lower-case (if a letter). A comma (,) causes the next character to be underlined. A dollar sign ($) causes the next letter or number to print as a superscript. C.2Q1F2.3 will print as C_2Q1F2_3, *i.e.*, the cipher for trifluoroethanol. Cards from Deck 1, or, more properly, the tapes made from these cards, will be used as a collection of compounds

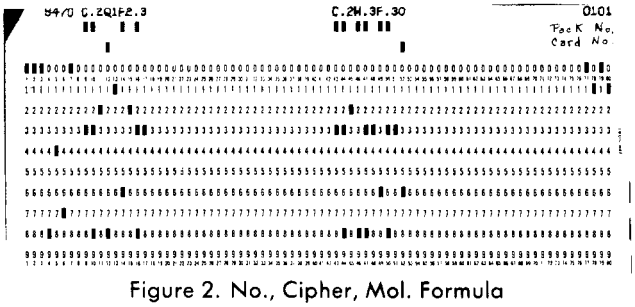

Figure 2. No., Cipher, Mol. Formula

*Presented before Division of Chemical Literature, American Chemical Society, St. Louis, Missouri, March, 1961.

**Chemical Abstracts Service, Columbus, Ohio.

to be scanned when a search for a given structure is desired.

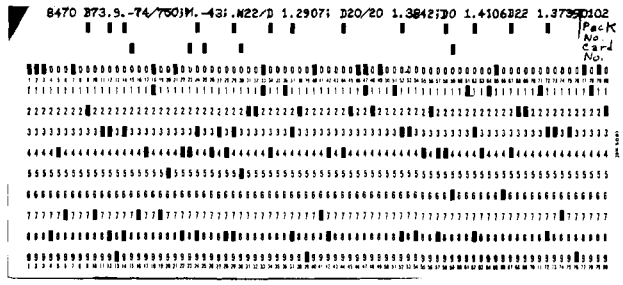

Figure 3. Physical Properties

Type 2 Deck-Physical Properties. Cards from the second deck (Fig. 3) bear, in the left-hand corner, the register number of the compound to be referenced. Following this "address," the physical properties of the compound are entered in the card. The first card made for any compound contains four pieces of data: (1) boiling point and the pressure at which it was determined, (2) melting point, (3) refractive index and the line and temperature used during the determination, (4) density or specific gravity and the temperature(s) at which it was determined. These data usually fill the first card; thereafter, as many more cards as are necessary are used to record all available physical constants for the compound in question.

One advantage of the IBM 1401 is that no fixed-field coding is necessary. The register number and the terminal signals (deck number and card number) are the only pieces of data that require an assigned position on the card. As long as the quoted constant is properly labeled (and standard abbreviations have been used as labels insofar as possible) the computer can retrieve it upon being given the label.

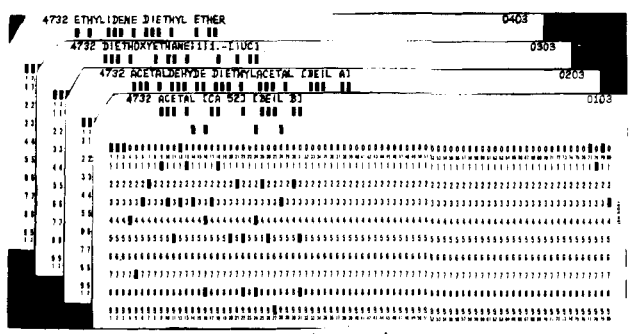

Figure 4. Nomenclature

Type 3 Deck-Nomenclature. Cards giving nomenclature of the compound in question (Fig. 4) bear the register number and the name (inverted) of the compound. If the name is an indexing name in *Chemical Abstracts*, it is followed (in brackets) by *C.A.* and a numeral for the volume number. If the name is an indexing name in Beilstein it is marked [Beil]; if it is a name known to be in I.U.P.A.C. lists, it is marked [IUC]. Originally, there was no intention of collecting this information, but already names have been found that are so far afield a mechanical device for relating them back to the structure has become essential. By means of this deck of cards, an unsystematic name or an ambiguous name can be related back to the cipher for structure.

Type 4 Deck-Literature Search. The summary card for each compound bears, up to the date of the latest index,

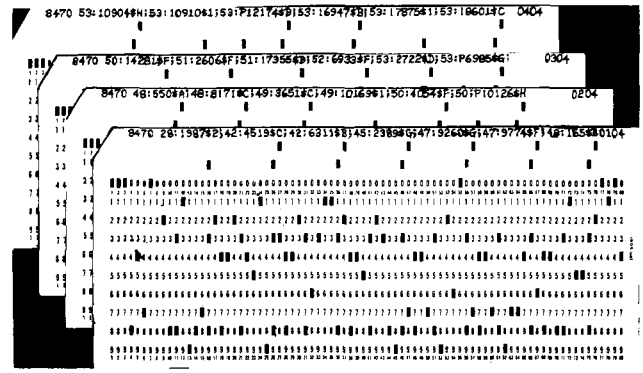

Figure 5. Literature Search

all the *C.A.* references to that compound, in order from most remote year to the present. Again beginning with the register number, as an address to signal the computer, search information is punched from the McBee summary card into an IBM card, then read into tape. Cards in this deck (Fig. 5) bear all the volume numbers, page numbers, and page area codes, on which information about the compound appears. Any one of the references or the entire list can be retrieved at will. A search can be limited to the most recent ten years, to the two years before the application date of a given patent, etc.

This card illustrates the real benefit from free-field coding. The operator successively punches volume number, colon, page number, the signal for a superscript, the area code, semicolon, then starts over again with the next volume number. There is no wasted space, and when the first card is filled she merely punches the deck number 04 and the card number, and goes on to the next card. One card will hold four or five references.

The usefulness of this deck is obvious. It is possible to screen the ciphers for structure (Deck 1) to locate compounds of a given type, then skip to the fourth type of card and print out all the references. In a relatively short time the whole story on a given class of compounds can be printed in conventional form or converted into new cards.

Type 5 Deck-Reaction Code. After considerable experimentation with the design of decks already described, there seemed still to be something lacking, namely, a means of showing the chemical properties, *viz.*, reactions, of the compounds in question. A mnemonic code which suggests the reaction has been devised and experiments with the code are being made. Some difficulty in referencing has been encountered. Among older compounds, many reactions predate *Chemical Abstracts*, and reference to the original literature becomes essential. But the plan of many older journals necessitates a whole series of numbers to pinpoint a reference. Means for abbreviating such information are being sought. When the reaction is recorded in *Chemical Abstracts*, citation is simple. Figure 6 shows a sample card from the Reaction deck; as is suggested by the mnemonic, Chl-01 is a chlorination, Diz-01 means a diazotization, Hyd-03 is basic hydrolysis, etc.; the numbers are volume and page references to *C.A.*

Type 6 Deck-Physiological Action. Dyson[1] has described the codification of data on physiological properties of various compounds, and this information will not be repeated here. Suffice to say, as in the case of prior examples, the register number of the compound is first

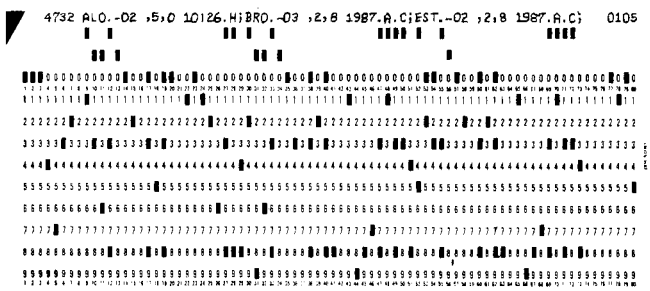

Figure 6. Reaction Code

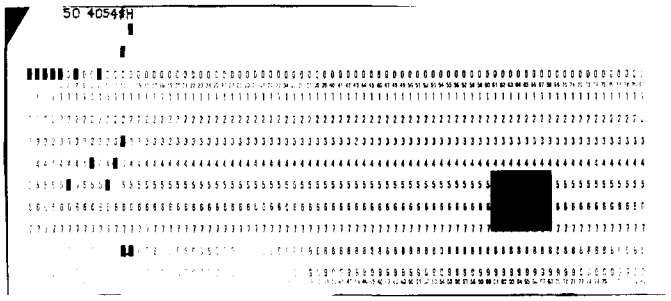

Figure 8. Abstract

punched into the card, thereafter, a code for the syndrome in question, the test animal, the route of administration, and the results obtained.

Type 7 Deck-Aperture Cards-Compounds. In addition to other data already coded, one further piece of information seemed desirable, *i.e.*, the graphic formula of each compound, correlated with the register number, cipher, and molecular formula. In a sense, the summary deck of McBee cards fills this need, but there are considerable advantages to machine sorting as compared with needle-sorting. To permit machine handling, Aperture Cards (Fig. 7) designed to withstand handling with an IBM 083 sorter are being prepared. The aperture card was designed so that the first part of the card from the initial deck can be duplicated onto the new card. Columns 61 through 69 of the aperture card are not available for punching; in that area a single frame of 16 mm. film is mechanically mounted. The insert is a diazo-process film, which resists scratching and survives sorting in mechanical equipment. The frame of film bears a photograph of the graphic formula of the compound. Other information, such

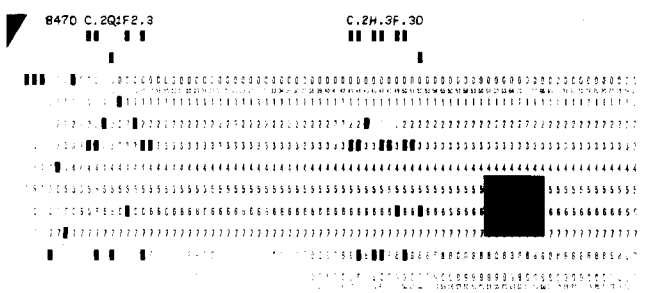

Figure 7. Graphic Formula

as bond angles and bond distances, can be recorded on the frame. Physical data too complicated for punching, *e.g.*, formulas containing many Greek letters, mathematical tables too long to punch, mathematical equations involving cumbersome fractions, can be recorded similarly. Information in the aperture card can be read, as is, with a simple reading machine, or it can be duplicated with a reader-printer; we are experimenting to see whether page-sized enlargement by Xerox is possible. An MMM Reader-Printer for strip film is also available, as well as another printer capable of duplicating the card, as is, to make another card.

Type 8 Deck-Aperture Card-Abstracts. Each useful deck seems to generate an idea for a new deck. The type 8 card was a direct result of handling the aperture cards just described. It was proved so convenient to have structural formulas available at the touch of a button, that the desirability of having abstracts similarly available became clear. Initially, the abstracts on fluorine compounds are being collected, with a view of simplifying the preparation of the fluorine lexicon. It has proved to be simple to do, because a complete set of references to fluorine compounds was already on IBM cards. All that remained to be done, to furnish the photographer with a list to work from, was to sort them with the Sorter. After placing them in order by volume and year, duplicates were eliminated. The photographer is now making and mounting film of the abstracts.

MACHINES AVAILABLE

Punching of the cards is done with a modified IBM 026 Keypunch. This machine is a sensing keypunch that is also used to operate the IBM 866 Typewriter with Document Writing Feature. Together, the equipment functions as a small, low-cost printer. As has been described earlier,[1,2] certain keyboard and operational changes were made in punch and typewriter, to permit of writing some special characters, subscripts and superscripts. It provides a cheap means of converting information in a deck of cards to a printed page or a continuous record sheet, and as a means of testing cards designed for the IBM 1401 (still undelivered). It can print 94 characters (including 10 numbers in 3 sets of positions). So far as the typewriter is concerned, a few more characters are possible, but the keypunch is the limiting factor—as yet there are no provisions to code cues for more characters. Although the little printer is slow (600 characters per minute) as compared with an IBM 1401 (420 lines per minute) it is still much faster than manual typing. Recently, the second of two book-length manuscripts was completed on the small printer. Aside from a few minor faults, it is highly satisfactory.

Two McBee 6103 Electric punches are also available. They are highly recommended over any type of manual punching of McBee Keysort cards.

None of these simpler machines require an air-conditioned room; however, Sorters, Accounting Machines, and other high-speed equipment do require air-conditioned and preferably humidity-controlled surroundings to reduce static build-up on moving cards, paper, and moving parts.

The air-conditioned and partially sound-proofed machine room in the Main Building contains five Keypunches and Verifiers, an IBM 083 Sorter (the most heavily used piece of equipment), a Collater, an Interpreter, a Reproducing Punch, and an IBM 407 Accounting Machine. Patent indexes are commonly made

with Sorter and Collater. Some of the simpler parts of *Chemical Titles* can be made on the IBM 407. However, the 407 is primarily an accounting machine and it cannot make logic choices. For the type of logic choices required, some time on a Computer is necessary.

The heart of the operational planning for literature searching and data retrieval is the IBM 1401 Computer. The 1401 was chosen because it offers a number of advantages. The operators are freed from the necessity of fixed-field coding. Ever since 1946, it has been possible to write ciphers unique for the structures of organic compounds, but equipment was not available which was capable of free-field operation and able to write lower-case letters, subscripts and superscripts. The 1401 fills both needs; it reads the entire card, from column 80 back to column 1, it provides 120 characters including lower-case letters, subscripts, superscripts, plus and minus signs, underlined numbers (essential to the cipher), and a number of special characters. It is no longer necessary to waste record-space by reserving fields for data not immediately available. It is no longer necessary to arrange data in columns, nor count spaces. It is no longer necessary to devise cumbersome codes for two-letter symbols of the elements; they can be written conventionally, with an upper-case and a lower-case letter. It is not necessary to segregate letters from numbers; the IBM 1401 can intersperse them at will. In addition, the 1401 is relatively easy to operate. It has a big memory (8000 characters) and more memory can be added, in successive 8000 character units.

From the searchers' point of view, the most exciting thing about the 1401 is the speed of the tape system. The large (10.5 inches in diameter) reels of magnetic tape (2400 feet) have a capacity of 16,012,800 characters in memory; in most records, about 14,000,000 are usable. Arithmetically, a reel equals about 175,000 cards; actually this is a little bit high because record design, and the kind of blocking maintained, influenced the capacity.

Preliminary counts, made in several ways from existing indexes, led to the conclusion that the number of known compounds is grossly underestimated. Probably the system must have capacity to handle 2.0 to 2.5 million compounds, of which 1.8 million are organic. Assuming that this estimate is correct, it will require only 10 reels to store the structures and molecular formulas of all organic compounds. Depending somewhat on the design of the record and the programming, and on the high-speed accessories available on the given 1401, the computer is potentially able to scan an entire reel in 3.5 to 4.0 minutes. At worst, it could be no longer than 10 minutes. This means that it is possible to scan all known organic compounds in 25 to 30 minutes, or at worst 100 minutes. Simple searches should take no longer than 30 minutes if the records are properly block-sorted. Moreover, preliminary indications are that three similar searches can probably be done at one time.

IMPLICATIONS OF THIS WORK

The aid of high-speed machines can now be available to the chemist in the laboratory. The research scientist can be freed from the repetitive tasks of assembling and organizing background material related to a project. This will make it easy to do a type of correlation that was previously possible only by hours and hours of chemist-time. For example, the limits and uses of a reaction can be explored by searching decks 1 and 5. The relation of structure and physical properties can be explored by searching decks 1 and 2. The effect of structure(s) on physiological action can be explored by searching decks 1 and 6. The relation of physical properties and physiological action can be derived from decks 2 and 6. All the indicative signs from such searches can point the way to new compounds and new uses of old compounds. Moreover, because of the unique properties of the cipher, and the mathematical relations inherent to it, there is now available a way to search for a given group or series of groups in complex structures and fused-ring systems. Finally, the entire information on any such search can be converted, by machine, to conventional typed pages. If necessary, it can be converted to a new tape or a new deck of cards.

BIBLIOGRAPHY

(1) G.M. Dyson, *J. Chem. Doc.*, 1, 24 (1961).
(2) G.M. Dyson and Elizabeth F. Riley, *Chem. and Eng. News*, 29, 72 (1961).

Experiments with the IBM-9900 and a Discussion of an Improved Comac as Suggested by these Experiments[1,2]

By MORTIMER TAUBE
Received August 24, 1961

This paper will present the system and machine considerations which led, in the first instance, to the development of the design of a Continuous Multiple Access Comparator; a description of the IBM 9900, which is IBM's reduction to practice of the COMAC principle; certain improvements in design of the COMAC to make it a more efficient low-priced device for the storage and retrieval of information; and, finally, the concept of using a wholly new method of logical comparison in order to achieve maximum efficiency from the use of E.D.P. equipment in the storage and retrieval of information.

In the literature on information storage and retrieval there is a generally recognized dichotomy between scanning systems and look-up systems, which have also

(1) This work was done for the Directorate of Mathematical Sciences, Air Force Office of Scientific Research, under Contract No. AF 49(638)-91.
(2) Presented before Division of Chemical Literature, American Chemical Society, St. Louis, Missouri, March, 1961.