LITERATURE CITED

(1) *The Indexer*, 9 Pensioners Court, Charterhouse, London E. C. 1.
(2) Maloney, C. J., "Practical Preparation of Internal Indexes," *Indexer* **5**, 81–90 (1966).
(3) Davenport, W. C., and J. T. Dickman, "Computer-based Composition at Chemical Abstracts Service," p. 9, paper presented before the Division of Chemical Literature, 151st Meeting, ACS, Pittsburgh, Pa., March 1966.
(4) Baxendale, Phyllis, "Content Analysis, Specification, and Control," in "Annual Review of Information Science and Technology," Carlos A. Cuadra, Ed., Vol. **I**, Chap. 4, Wiley, New York, 1966.
(5) Newman, S. M., "Storage and Retrieval of Contents of Technical Literature, Non-Chemical Information," 2nd Supplementary Report, U.S. Patent Office, Washington, D. C., 1958.
(6) Yngve, V. H., "The Feasibility of Machine Searching of English Texts," International Conference on Scientific Information, National Academy of Sciences, Washington, D. C., 1959.
(7) Sedelow, S., and W. Sedelow, "Stylistic Analysis," in "Storage and Retrieval of Contents of Technical Literature, Non-Chemical Information," *op. cit.*
(8) Walker, J. F., and R. F. Schirmer, "The Indexing of Technical Books," J. CHEM. DOC. **6**, 26–30 (1966).
(9) "Annual Review of Information Science and Technology," *op. cit.* p. 389.
(10) Maloney, C. J., and M. N. Epstein, "Progress in Internal Indexing," *Proc. Am. Document. Inst. Ann. Meeting* **1966**, pp. 57-62.
(11) Bondi, A., "On Error Prevention," J. CHEM. DOC. **6**, 137–142 (1966).
(12) Hines, T. C., and J. L. Harris, "Computer Filing of Index, Bibliographic, and Catalog Entries," p. 126, Bro-Dart Foundation, Newark, N. J., 1966.
(13) Artandi, Susan, "Book Indexing by Computer," Ph.D. thesis, p. 207, Rutgers—The State University, New Brunswick, N. J., 1963.
(14) Borko, Harold, Ed., "Automatic Language Processing," Wiley, New York, in press.
(15) Maloney, C. J., James Dukes, and Sterling Green, "Indexing Reports by Computer," in "Technical Preconditions for Retrieval Center Operations," Benjamin Cheydleur, Ed., p. 13-28, Spartan Books, Washington, D. C., 1965.
(16) Maloney, C. J., "Semantic Information," *Am. Document.* **13**, 276–287 (1962).
(17) Glickert, Peter, "A Codification of English Words," The Author, p. P5, 1966.
(18) Olney, J. C., and D. L. Londe, "Language Processing. Anaphoric and Discourse Analysis," p. 4, unpublished paper.
(19) Clarke, D. C., and R. E. Wall, "An Economical Program for Limited Parsing of English," AFIPS Conference Proceedings, **27**, Spartan Books, Washington, D. C., p. 307–316, 1965.
(20) Beveridge, Gerald, and C. J. Maloney, "The Biological Laboratories Information Retrieval Program," p. 99, available from Office of Technical Services (AD 277 544), June 1962.

Keyboarding Chemical Information*

R. G. HEFNER, P. M. KEESECKER, and D. F. RULE
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received September 21, 1967

During 1967, the Chemical Abstracts Service (CAS) will print some 570 million characters in its publications. A computer-supported keyboarding system has been developed at CAS to handle the data input process. The varied character set, approximately 1500 different type pieces, is being entered into the system through standard keypunch and typewriter keyboards. Input conventions have been developed following extensive analysis of character frequency counts. Substantial use is made of input formatting, shortcut techniques, and recognition of routine grammatical construction and of the natural structure of the data in providing inbuilt signals to the computer to allow automatic case and face changes. These techniques supplemented by a simplified keyed-flagging and a mnemonic code system are described in this paper.

The printed information services offered by Chemical Abstracts Service will, in 1967, comprise 131,000 pages carrying 570 million characters; the present rate of chemical information growth indicates that by 1970 CAS will publish 174,000 printed pages containing 760 million characters exclusive of any new services. To provide chemists and chemical engineers with ready access to this store of information, CAS is converting to computer base and will make information stored in computer files available not only for producing printed information tools, but also for direct searching. This conversion, which has recently been described in papers by Davenport (*1*) and Tate (*2*), requires that the most efficient means be developed for translating chemical information into machine language by keyboard devices assisted by appropriate computer programs.

Some 1500 different type pieces, or characters, are required to represent the nonstructural material in

*Presented in part before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.

Chemical Abstracts. These include italic and boldface characters, small capitals, Greek and other special alphabets, and a variety of special symbols, as well as the standard body text. Several additional symbols are required for diagrammatic material such as structural diagrams. To handle these many characters, publishing operations have traditionally relied upon equipment through which a keyboard operator controls a relatively large set of characters, while special characters are hand-set piece-by-piece as necessary. Standard office machines such as the typewriter do not offer this flexibility; such machines have fewer than 100 different characters and there is only limited opportunity for the introduction of special characters. Therefore, for input to a traditional publishing operation, copy produced on office keyboarding equipment must be hand-marked to indicate special characters such as Greek letters or small capitals, and special treatment for letters such as italics or boldface.

Viewed in the light of traditional publishing operations, the standard office keyboarding equipment is poorly equipped for keyboarding chemical information. However, viewed in light of computer-based systems, the problem takes on a different aspect. Because the computer can manipulate data, rearranging it and reformatting it by instruction, the appearance of the data at input need not match the output appearance so that the input task can be designed separately from the output requirements. This has several important consequences for keyboarding. In the first place, one can distinguish between *different* and *unique* letters. A ten-point, uppercase, Baskerville, italic d and an 8-point lowercase, Spartan Bold d can be considered *different* versions of the same *unique* character —the fourth letter of the alphabet. Thus uppercase, italic, boldface, small capital, superscript, and other versions of a character can all be considered different modifications of the same unique character. Viewed in this manner, the 1500 different characters required for chemical information reduce to about 125 unique characters.

Of course, 125 characters are more than the normal office machine can represent. However, at this point, another factor enters the picture: the frequency distribution of characters in running text shows that several characters are used very rarely. A statistical study on a sample of text from three issues of *Chemical Abstracts* Volume 63 shows the frequency of use of unique characters in *CA* rises very sharply between 36 and 50 characters and then levels off into a very long "tail." Figure 1, which shows the frequency-distribution curve, indicates that 94.50% of the material found in running text can be represented with 36 characters, assuming that suitable "flags" are available to indicate uppercase, italic, boldface, superscripts, subscripts, and other versions of each character. These flags are themselves nothing more than special characters, and they are included among the 36 characters. With 64 characters, the number available on the standard keypunch, 99.59% of the volume of text can be represented, and with 88 characters, the number available on the standard typewriter considering up-shift and down-shift, 99.95% of the volume can be represented.

The preceding discussion shows that in a computer-based system, the requirement for 1500 different type pieces to represent chemical information can be reduced

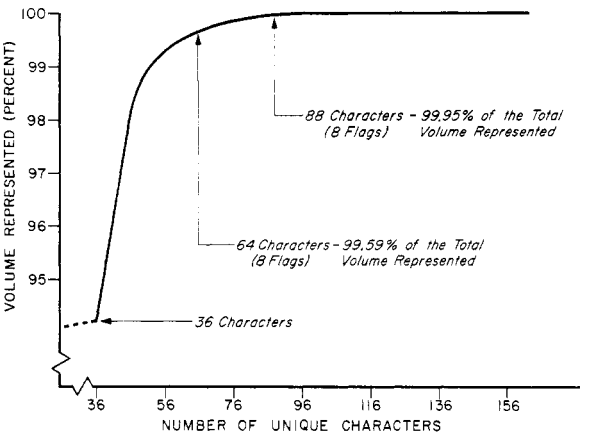

Figure 1. Number of unique characters required to represent a given percentage of chemical text

to a task appropriate for standard office keyboarding equipment by applying the concept of unique characters rather than different characters and by applying the character frequency distribution in running chemical text. Moreover, the difference between a keypunch, with 64 characters, and the typewriter, with 88 characters, is not significant in terms of their ability to handle chemical information. Thus, a choice between these two different types of equipment must be made on other factors, as discussed below.

CAS EXPANDED CHARACTER SET AND SPECIAL KEYBOARD

To take advantage of the distinction between unique and different characters, CAS has developed an Expanded Character Set for use in the computer. This set provides for 256 unique characters, each of which may have 98 variations—a potential total of more than 25,000 different characters. Characters input to the computer are translated to the Expanded Character Set for storage, and material output from the computer file is translated from the Expanded Character Set to whatever character set is available on the output device being used. Thus, point size and font are determined at output time, not at input time. Moreover, the CAS Expanded Character Set can be used with many different types of input and output devices and easily be converted to character sets of lesser detail such as the American Standard Code for Information Interchange (*3*).

To take advantage of the flexibility provided by computer-oriented keyboarding, CAS has designed special keyboarding conventions for use with the keypunch and the standard typewriter. In essence, these conventions create a special keyboard for use with chemical information. The typewriter keyboard is illustrated in Figure 2; the following discussion will deal with this keyboard, although the keypunch conventions are very similar. The character set illustrated in Figure 2 is available from Camwil Corp.(*4*) as a type element for the IBM selectric typewriter and other devices using the selectric mechanism.

The characters on the keyboard represent those that occur frequently in chemical text. The keyboard contains

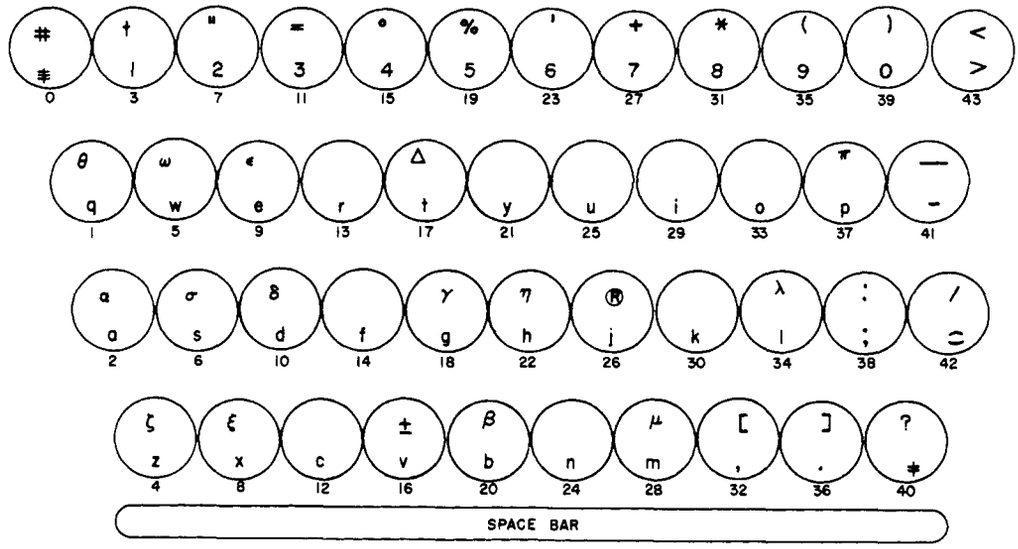

Figure 2. Keyboard layout for special CAS keyboard

the lowercase Roman alphabet, the frequently used Greek letters, the numerals, punctuation, and many special characters. Several of the special characters are "flags" that signal the computer to take special action on following characters. For example, key 40, the double dagger (‡), is used as a flag to obtain capital letters; whenever the flag precedes a lowercase Roman letter, that letter will be capitalized in the computer record.

The advantages of this flagging technique should not be underestimated. The use of a special symbol to signal that the following letter is to be capitalized makes it unnecessary to carry the uppercase Roman alphabet on the keyboard. Thus, 26 characters are freed for use as special symbols. This illustration highlights the fact that some special flags available on the keyboard resemble normal capital letters. However, these letters are *never* used as capitals in text; they are always used as flags.

Since the keyboard eliminates all but the most frequently occurring characters, the keys for the additional coding needed to identify character variations and characters not included directly on the keyboard become available. The use of a flag for capital letters has already been indicated. Another special flag used on the keyboard is key 0, triple dagger (≡), that signals the use of three-characters mnemonic codes for special characters that occur infrequently in *Chemical Abstracts* printed text. For example, the infinity symbol is represented by the code: ≡ inf. Similarly, the dollar sign, which occurs rarely in chemical text, is represented as: ≡ dol. This system is flexible in that a virtually unlimited set of codes can be defined and added as the need arises.

These illustrations show how a standard office typewriter keyboard can be given greatly expanded capabilities through programmed computer assistance and through relatively simple coding conventions. These conventions are fairly easy for a typist to learn, because the keyboard still strongly resembles a standard typewriter keyboard. However, while important progress has been made in optimizing keyboard design for CAS input tasks, this work is not yet complete. We are in the process of gathering additional character-frequency data based on ongoing keyboarding at CAS. These data will be analyzed to determine the optimum positioning for the various flags and special characters used. Of particular importance will be the assignment of upshift positions, since each character in an upshift position requires one extra keystroke to type.

REDUCING KEYSTROKES THROUGH COMPUTER ASSISTANCE

Computer assistance for keyboarding permits not only improved character-by-character input; it offers important capabilities for reducing the number of keystrokes required for input. Special input conventions such as the use of abbreviations for long or commonly occurring words can be designed into the keyboarding task and the computer can be programmed to interpret these conventions—*e.g.*, to expand the abbreviations—as needed.

An example of program assistance is a convention that saves keystrokes in the keyboarding of molecular formulas. Using a single-character flag to indicate that what follows is a molecular formula, the keyboarder types the formula without capital letters, and without subscript numerals. Thus, a typical molecular formula would be keyboarded as: c16h16n4o5s2. When the computer receives this information, a program supplies the correct format to print the formula as it normally appears: $C_{16}H_{16}N_4O_5S_2$.

Another type of shortcut that has been programmed into the CAS computer system saves time for both chemists and keyboarders by automatically expanding the common letter combinations used to signify functional groups in chemical nomenclature. The use of such frequently occurring abbreviations as "me" for methyl, "ph" for phenyl, and "pr" for propyl can be carried through the entire input process from chemist to keyboarder, while the computer will expand the shortcuts into the full chemical term. Thus, the name keyboarded as t-bu acetate will be recorded as *tert*-butyl acetate. Not only does this save time and keystrokes, but it also assures a consistent computer record. That is, a name can be input at two different times or by two different people, one using the shortcut and one not using the shortcut, but the resulting computer records will be the same in both cases.

Another way to streamline the keyboarding task is to program the computer to supply characters based on the context of the information. An example is the placing of italics and capital letters in names of chemical compounds. The structure of chemical nomenclature lends itself to programmed algorithms for reviewing and automatically supplying italics and capitals in systematic chemical names. To reduce the number of keystrokes that must be recorded for chemical names, CAS has developed and programmed rules for the machine insertion of capitals and italics, expansions of certain alphabetic strings, and editing of some punctuation characters. For example, there is a group of 49 character strings such as "cis," "trans," and "erythro" that will be automatically italicized when they are set off by punctuation in a compound name. Similarly, another computer routine automatically capitalizes the first letter of a name, disregarding prefixes and single-letter locants. Because the computer supplies italics and capitals, it is unnecessary to enter them via the keyboard. These automatic editing procedures are described more fully by Park *et al.* (*5*).

EVALUATION OF KEYBOARDING EQUIPMENT

The techniques described above for improving the efficiency of the keyboarding task are for the most part independent of the particular type of keyboard device used. A standard office keypunch or typewriter offers enough keys to permit the coded input of chemical information, and the coding conventions are similar for each type of machine. Therefore, the choice of a keyboarding device for the input of chemical information to a computer-based system depends primarily on the costs of operation rather than on the technical capabilities of the machine. Specifically, the criteria CAS has used for evaluating keyboarding devices are:

The ability of the machine to reliably produce computer-readable data.

The ability of the machine to produce direct "hard copy" where hard copy is required.

The cost of operating the equipment, including both labor and machine cost.

The cost of transferring data from the format output by the keyboard device (*e.g.*, punched cards) to the format required by the computer (*e.g.*, magnetic tape).

The ease of operator training and the degree of proficiency that an operator can reach.

Willingness and ability of the equipment supplier to modify his equipment where desirable (*e.g.*, to provide a customized character set).

CAS has recently applied these criteria in a series of tests involving four commercially available keyboarding devices. These devices differ in the number of keys offered on the keyboard, and the medium on which the data are recorded (*e.g.*, paper tape, punched cards, or magnetic tape), and their ability to produce "hard copy", and in several other characteristics.

The results of these tests apply to the specific situation and requirements of the CAS keyboarding operation. Therefore, the results should not be generalized to other institutions with other requirements.

The IBM 029 Keypunch (*6*). The keypunch has 64 characters, produces punched cards, but no hard copy. At CAS, our experience has been that this device can be rented at the lowest cost of all tested, and is the most reliable mechanically. However, in our application, many keyboarding tasks require a special coding of characters at input time. As operators must learn the coding conventions, a relatively long training time is required for them to reach proficiency; even then, the frequency of operator-induced typographical errors is higher than with some of the other devices tested. In addition, the punched cards produced by the keypunch must be converted to magnetic tape for use at the CAS computer installation, thus incurring cost for the cards and costs for conversion to magnetic tape.

The Mohawk 1101 Magnetic Tape Data Recorder (*7*). This device has 64 characters on a standard keypunch keyboard. It records data directly on magnetic tape, thus obviating the need for punched cards or a conversion step from cards to magnetic tape. An additional benefit is that this equipment performs duplicating and skipping procedures much faster than a keypunch. However, the Mohawk 1101 produces no hard copy and its operator training requirements and error potential are similar to those for the IBM 029 keypunch.

The Dura Mach 10 Typewriter (*8*). This typewriter records data in two forms—in a punched paper tape for machine input, and as hard copy for editing. At CAS, operator training time is relatively short with this device, even with the special CAS keyboard (Figure 2). However, operating the paper-tape-typewriter in the CAS system requires an extra processing step, the transfer of data from paper to magnetic tape for computer use. CAS has also experienced mechanically induced errors involving the tape punch.

The Mohawk 1181 Magnetic Tape Typewriter (*7*). This device records directly onto magnetic tape and simultaneously produces hard copy. The keyboard is the special CAS keyboard, and operator training time is relatively short, CAS has found the Mohawk 1181 to be mechanically reliable.

As a result of our testing over the past two years, we expect to convert all CAS production keyboarding operations to one or the other of the magnetic-tape-recording devices, depending on the need for hard copy in the particular operation.

THE KEYBOARDING OF CHEMICAL STRUCTURES

A very specialized keyboarding task in the handling of chemical information is the keyboarding of chemical structural diagrams. The mechanization of chemical information systems often requires the development of a computer-based system for handling chemical structural information. As computers are not, at present, commonly assigned the task of handling pictorial representations of structural diagrams, such systems require that some method be found to translate from the structural diagram as drawn by the chemist into a form of easily handled computer coding.

CAS is now operating a large-scale Chemical Compound Registry System which directly handles structural diagrams. In the early stages of Registry System operation, CAS used an input method for structural data in which the two-dimensional diagram drawn by a chemist was coded by a clerk and converted to a "connection table" input to the computer as a string of numbers and letters representing atoms and bonds. Because of the large scale

of our operations—approximately 700,000 different structural diagrams have been processed to date—we have investigated a number of alternative input methods for structural data.

One such input method, which was initially developed at Walter Reed Army Institute of Research (*9*), is the direct typing of structural diagrams, with the computer performing the work of converting the structure to a numerical format for filing. Most of our work has employed commercial typewriters, modified according to the requirements defined by J. Mullen at the Shell Development Co. (*10*). CAS is indebted to Shell for help in the mechanical modification of the commercially available Dura Mach 10 tape-generating typewriter to meet structure-typing needs. Computer input from this device requires the transfer of data from the punched paper tape to computer-readable magnetic tape. In related work, CAS, the Mohawk Data Sciences Corp., and the Invac Co. (*11*) have cooperated in developing a structure typewriter that records data directly on magnetic tape in an operation similar to the operation of the Mohawk 1181 Data Recorder.

CAS has developed the structure-typing conventions and the computer programs necessary to permit structures typed with either of these devices to be input directly to the Registry System. That is, the record produced in structure typing is directly translated into a connection table and registered. In the past several months, more than half of routine registration has entered the computer through the structure typing keyboard. Because the typist is copying a hand-drawn structure rather than translating it into a numerical code, there are fewer input errors. And, although the economic data are still developing, there is an early indication that structure can be typed at least as fast as connection tables can be generated and keyboarded.

TOWARD A COMPUTER-BASED COMPOSITION SYSTEM

Keyboarding tasks will enter heavily into the future of CAS services. Eventually all CAS publications and information services will be produced through a composition system operated by the computer; this system will depend on a single intellectual manipulation of the information and a single keyboarding of the data input to the system. The information will be permanently recorded in machine-searchable form for immediate and future reuse for both computer searches and publications. These permanent files can be searched for specific information at CAS or the information from the files can be duplicated on tapes purchased by a customer and sent to him for searching at his own location. The end result will be a permanent, reliable, and useful machine language record of digested and indexed chemical knowledge and a system that can function with great speed and be capable of intermixing both text and graphic representation in a continuous operation. The design of effective, efficient keyboarding techniques will be central to this system.

ACKNOWLEDGMENT

Portions of this work were supported by the National Science Foundation, the National Institutes of Health, and the Department of Defense through contract NSF-C414.

LITERATURE CITED

(1) Davenport, W. C., "A Complete System for Handling Chemical and Chemical Engineering Information," FID Meeting, Rome, June 1967.
(2) Tate, F. A., "Progress Toward a Computer-Based Chemical Information System," *Chem. Eng. News* **45**, No. 4, 78–88, 90 (1967).
(3) American Standard [Code for Information Interchange], Standards Institute, United States of America, New York, 1965.
(4) Camwil, Inc., 835 Keeaumoku Street, Honolulu, Hawaii (Part Number 127M).
(5) Park, M. K., K. Kenny, and P. E. Swartzentruber, "Computer Editing of Chemical Nomenclature," Division of Chemical Literature, 154th Meeting, ACS, Chicago, Ill., September 1967.
(6) International Business Machines Corp., 618 S. Michigan Avenue, Chicago, Ill.
(7) Mohawk Data Sciences Corp., Harter Street, Herkimer, N. Y.
(8) Dura Business Machines, 32200 Stephenson Highway, Madison Heights, Mich.
(9) Feldman, Alfred, "A Proposed Improvement in the Printing of Chemical Structures Which Results in Their Complete Computer Codes," *Am. Document.* **15**, 205–9 (1964).
(10) Mullen, J. M., "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures," J. CHEM. DOC. **7**, 88–93 (1967).
(11) INVAC, 26 Fox Road, Waltham, Mass.