(27) "General Proposal, Library Information Transfer System," Alden Electronic and Impulse System Recording Co., Inc., Westboro, Mass., 1967.

(28) "Alden Dial A Document—Mail by Telephone for Information Storage and Retrieval Centers," Alden Electronic and Impulse Recording Equipment Co., Inc., Westboro, Mass., 1969.

(29) Sophar, G. J., and L. B. Heilprin, "The Determination of Legal Facts and Economic Guideposts with Respect to the Dissemination of Scientific and Educational Information as It Is Affected by Copyright," Committee to Investigate Copyright Problems Affecting Communication in Science and Education, Inc., Washington, D. C., 1967. **PB 178,463.** ERIC No. **ED 014621.**

(30) Weil, B. H., and J. H. Kuney, "ACS Developing Policy on Copyright," *Chem. Eng. News* 48(35), 37-9 (1970).

(31) Immergut, E. H., ACS Copyright Committee communication, March 24, 1971.

(32) "GPO Begins Arduous Task of Defining Needs and Impact of Proposed Micropublishing Program," *Micrographics News & Views* 2(4), 1 (February 28, 1971).

(33) Taylor, H., "Magnavox to Recommend Fax Compatibility Ruling," *Electronic News* **15,** 32 (August 24, 1970).

(34) Gardner, A. E., "The Future Isn't What It Used to Be," *Publisher's Weekly* **199**(2), 42-5 (1971).

(35) Mezrich, R., "A Medium for the Message: The Most Exciting New Medium for Storing and Retrieving Images Is the Hologram." *Industrial Research* **11**(11), 58-60 (1969).

(36) "Video Records: Everyone's Future Toy," *Publisher's Weekly* **199**(1), 30-2 (1971).

(37) Chartrand, R. L., "The Fourth Generation: Intimations of Reality," *J. Data Management* **8,** 99-102 (September 1970).

(38) Avedon, D. M., "Transmission of Information—New Networks," *Special Libraries* **61,** 115-18 (1970).

(39) "British Push R&D on Fiber Telcom Link," *Industrial Research* **13**(1), 33, 34 (1971).

(40) "How to Copy a Video Display," *Business Week,* No. **2139,** 30 (August 29, 1970).

The What and How of Computers for Chemical Information Systems*

HERMAN SKOLNIK
Hercules Incorporated
Research Center
Wilmington, Delaware 19899

Received March 1, 1971

What literature chemists need to understand and to do to use computers efficiently and productively is discussed in terms of when a computer is needed, which computer, what a computer is and how it works, computer logic, and designing a chemical information system for computer processing.

The computer has a long past but a very short history. Its antecedents are the abacus of the ancient orient and various calculating devices of ancient Greece, the mechanical digital calculator made by Blaise Pascal in the 17th century, and the mechanical calculator designed by Charles Babbage early in the 19th century and from which evolved the electromechanical accounting machines of the early 20th century. The real beginning of the history of the computer was probably in the late 1930's when Howard H. Aiken of Harvard University and George R. Stibitz of Bell Telephone Laboratories introduced automatic calculators with relays. This was closely followed by the work of J. Presper Eckert and John W. Mauchly, at the University of Pennsylvania, during the early 1940's which resulted in ENIAC, the first electronic computer.

Computers of the 1940's and early 1950's were vacuum tube based. For example, ENIAC contained 18,000 vacuum tubes. These vacuum tube computers are now called first generation computers. They performed arithmetic (addition, subtraction, multiplication, etc.) and logical operations (comparison, selection, and rejection) in a thousandth of a second or less. Magnetic tapes were the primary mass storage devices and, consequently, data could be processed only sequentially and by batching techniques, very much like, but faster than, the electric accounting machines of the 1930's and 1940's.

The first generation computers evolved into what are now called second generation computers. These computers of the 1950's and early 1960's used magnetic core working storage with transistors and diodes replacing the vacuum tubes of the first generation computers. They performed the same arithmetic and logical operations as the first generation computers, but at speeds of microseconds, or millionths of a second, and with the added advantage of direct access to mass storage—i.e., many thousands of stored records, any one of which could be retrieved in a fractional part of a second.

During the 1960's, the third generation computer was introduced with speeds of nanoseconds or billionths of a second. These computers were characterized by integrated microcircuitry.

At the beginning of 1970, the number of computer installations—viz., second and third generations—totaled about 70,000 in the United States. This growth of the computer industry over the past 25 years is without equal among technological developments. Its glamor and impact have been unique, by any standard, and today computers are a pervasive part of daily living of practically everyone in the United States.

Computers are very much present in the chemical industry for mathematical and accounting applications. I know of no chemical company which has a technical library or a

*Presented before the Division of Chemical Literature, 161st Meeting, ACS, Los Angeles, Calif., March 29, 1971.
Hercules Research Center Contribution No. 1552.

technical information group and which is without a second or third generation computer. Yet only a small number, a very small number, of these groups has applied computers for chemical information systems.

The purpose of this paper is to delineate what literature chemists need to understand and do to use a computer effectively and efficiently.

WHO NEEDS A COMPUTER?

The computer is a thing of beauty. It is a precise and exacting machine. But it can and undoubtedly will be the most expensive item in your budget. And it can and undoubtedly will be the most frustrating and emotionally upsetting mechanism you have ever encountered. Although the computer is fraught with numerous intellectual and economic problems, when used effectively and efficiently it yields results and products not possible physically or economically by any other mechanism.

There is no formula that we can use to determine the need for a computer. I think there are, however, two important principles that can serve as guide lines:

> The use of a computer to process and print out information for a chemical information system should result in a net cost saving over that of the system before computerization.
>
> The use of a computer to process and print out information for a chemical information system should result in a better product for the information user and in products not economically or easily obtainable by other methods.

Once stated, these two principles appear to be obvious. Yet relatively few computerized information systems follow these two principles. There are several reasons why so many users do not realize the computer's potential. There is a strong tendency to preserve an information system's traditional form in the computerized system. We thus tend to be concerned with how to do in the computer what was always done, doing the same tasks with the same objectives as they were before using the computer.

The most important thing that we must know about the computer is that it does not think. The computer, however, is extremely and entirely logical, and the most important benefit of this realization is that it makes us think. This thinking can be as productive as using the computer.

Information systems in the past were relatively simple and straightforward because they were unit operations for single objectives. This was true whether we used 3 × 5 card files, visible card files, edge notched punch cards, uniterm posting cards, optical coincident cards, or tab cards with sorters, collators, and tabulators.

WHY COMPUTERIZE?

Basic to any information system is the recognition and understanding of the information needs of the users and potential users. Without this recognition and understanding, an information system is but an exercise in fiction; with recognition and understanding, an information system is an intelligent and intellectual accomplishment that saves considerable time, energy, and costs for the users.

Additional savings of time, energy, and costs are possible with well-conceived computerized information systems. On the other hand, computerization is not indicated if these savings cannot be realized.

Let us consider a journal literature system based on the contents of 520 journal titles for a user group of 200 chemists who require both an awareness and retrieval tool. In this document base of 520 journals, let us assume an average of ten issues per journal per year, or a total of 5200 journal issues per year, and an average of 20 articles per issue, or a total of 104,000 articles per year. Let us further assume the almost improbable that the flow of journals from the publishers and through the mails is steady—i.e., we receive 100 journal issues per week containing a total of 2000 articles. But in terms of the users, the number of articles among these 2000 which are pertinent to their spectrum of information needs probably would be fewer than 200, or less than 10% of the total. An information system that inputs and outputs the total document base, consequently, would be at least ten times larger than required by the information needs of the users. Furthermore, the articles pertinent to the users would be diluted at least tenfold by articles not pertinent.

Dilution is an important economic factor in both awareness and retrieval tools. For example, an awareness tool that adds one hour of unnecessary reading time per week for 200 chemists reduces over-all productivity by five man-years per year. More critical, however, is that excessive dilution discourages use of an awareness tool. Retrieval tools, similarly, can be vitiated by excessive dilution.

The above document base is an excellent candidate for computerization providing one input yields many products. The number of products, however, depends on the needs of the users. Let us assume that the users require an awareness tool and a retrieval tool from the following viewpoints:

1. Subject
2. Author
3. Location of author (university, company, and government agency)
4. Class
   a. Discipline of science—e.g., analytical, organic, polymer, chemical engineering, etc.
   b. R. and D program or project—e.g., terpene chemistry
   c. Uses or applications of chemicals
5. Journal

The awareness tool and each of the retrieval tools can be products from the same input. Furthermore, because a journal-based information system exists, other needs can be tied into it, such as journal subscriptions, routine circulation of journals, charge-out of journals, and statistics on journal use. Designing the computerized information system will be discussed later.

An information system is a potential candidate for computerization when it requires duplication of input and multiple files of output. In the above document base, a 3 × 5 card file for only the subject index and author index could require an average of four subject cards and two author cards (with abstracts) per article entered into the system, or a total of 1200 cards per week (for 200 articles per week). Typing and filing 1200 cards (with an average of seven lines per card) per week is a full time job for a good clerk typist. Keypunching seven tab cards per article for input into the computerized information system (subjects, abstract, author(s), reference, author's location, and classification codes) for 200 articles per week can be done within eight hours per week by an average keypuncher. The computer is programmed to do the rest.

In summary, computerization of an information system is indicated and feasible when some or all of the following can be realized:

1. Reduced input costs (typing vs. keypunching)
2. Elimination of filing costs
3. Many products from one input

4. The computer data base is used for unrelated tasks (for example, in the above, awareness, retrieval, journal subscriptions)

WHICH COMPUTER?

Computer makers have yet to design and build a machine that is adequately oriented to the tasks and problems of the literature chemist. But, if they had, it is unlikely that enough information groups in the chemical industry could afford one. Rental costs of computers are high relative to the total budget of most information groups, and only a relatively few information groups in the chemical industry could keep a reasonably adequate computer busy for more than one or two hours a day. Consequently, the question in the world of chemical documentation is not: "which computer is best?". But "how can we use profitably the computer or computers already within the company for accounting and mathematical operations?". For the past several years, we have had a second option: the use of outside computer service establishments, either on-line or off-line, and this second option also warrants investigation.

It is important to remember, however, that chemical documentation tasks and problems are similar to those of accounting and inventory tasks in requiring massive inputs and outputs with a relatively low demand on computer core processing time. Mathematical operations, on the other hand, require a negligible input and output but with a high demand on the computer core processing time.

In tying into a computer already in the organization, we need to establish first whether the available computer is adequate for our tasks and operations, and second whether software is available with the computer so that we can use the computer profitably without becoming overly involved with system analysis, the most expensive part of using a computer.

Until the late 1960's, there were essentially two kinds of computers: scientific and commercial, each with a specialized programming language. With the advent of multipurpose or general purpose computers, by IBM, RCA, CDC, Burroughs, Honeywell, NCR, and others, this difference disappeared. Also, with the introduction of general purpose computers with very large core storage the way was opened for time-sharing of a centralized computer system by many functions at many locations within an industrial organization.

Time-sharing is the system by which a single computer is used by many users through a variety of input and output devices. Input and output devices are relatively slow in comparison to the speed that the computer can process data. It is therefore economical to have an optimum number of users interacting with the computer through the slower input and output devices. The advantage to each user is that time-sharing makes available to him a powerful and costly computer at a relatively low cost.

The important point is that using or sharing a computer efficiently and effectively requires that the user be as knowledgeable as possible of what a computer is and how it works.

WHAT IS A COMPUTER AND HOW DOES IT WORK?

A computer is an electronic machine designed to perform logical and arithmetical operations by a set of circuits interacting with a set of instructions. What a computer does can be done by anyone who knows a common numbering system, elementary arithmetic, and the ABC's, providing he is willing to follow instructions explicitly without question and without deviation.

Until relatively recently, a computer consisted of essentially one black box, the central processing unit (CPU). The heart of the CPU is the command repertoire which is unique for each specific type of computer and whose electronic circuitry is dedicated to controlling arithmetic and logical operations. Most of the CPU is available for data that are being processed and instructions for processing the data.

To operate on data for desired results requires units for input data and for output results. Thus, in addition to the CPU, peripheral units—such as auxiliary memory, card readers, disks, and tape units for input and card punches, disks, tape units, and printers for output—are components of the computer configuration. Many computers today are disk oriented systems (DOS), the disk unit being essentially the second black box in the computer configuration shown in Figure 1.

Software consists of the total programs and routines associated with each unique type of computer and as usually supplied by the manufacturer. A computer without software is like a newborn man—i.e., born as an adult with no education and no experience. Just as education and experience enables a man to function, software enables a computer to function, and the greater the sophistication (education and experience) of the software, the more powerful the computer and the higher the level of language that we can use to communicate with the computer.

The assembler and compiler consume a large portion of the software; job control language is another relatively large unit. Although not relatively large, the arithmetic and logic unit is a very important part of the software. Other software units may be application-oriented programs, time-sharing, text editing, emulation, etc.

Machine language, which is a binary numeric language without grammar and sentence structure, is the most direct and efficient way to communicate instructions to a computer. Although the final format of programming instructions is in machine language, only a very few users of computers have had the time, energy, and desire to become familiar and expert with it. Assembly language is a symbolic language which uses symbolic and mnemonic programming aids; the assembler converts assembly into machine language. The compiler converts higher level languages, such as FORTRAN, COBOL, PL/1, etc., which are machine-independent, into machine language. The difference in computer efficiency between assembly, a machine dependent language, and FORTRAN, a machine-independent language, is apparent from the fact that if a program in assembly requires a given unit core—e.g., 10K—a program in FORTRAN will take at least twice, but more likely two to five times, as much—i.e., 20-50K. But the ease in learning and using a higher level language more than offsets the difference in computer efficiency.

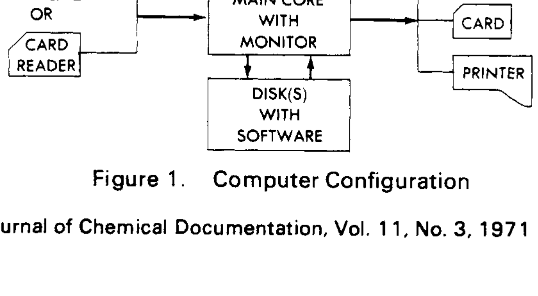

Figure 1. Computer Configuration

The program punched into tab cards, one instruction per card, is known as the "source program." The converted program in machine language from the assembler or compiler is known as the "object program" and is what the computer uses to manipulate the data.

Because users of computers desired to direct the functioning of the computer's operating system, such as loading and executing programs and outputting results on specified devices, the job control language or cards and its software were introduced. A deck of control cards is primarily a mechanism for introducing and identifying a job, a job step, the program or procedure to be executed; linking data sets to the program; putting limits on the data in the input stream, and indicating the end of the job.

Keypunching is still the primary method for delivering data, source programs, and job control statements to the computer. Keypunching is exceedingly slow and fallible in comparison to other computer components. Although there has been considerable activity in attempts to replace keypunching, the IBM 029 keypunch dominates this phase with key-to-tape gaining slow acceptance. In addition to being slow, slightly less than normal typing rates, keypunching's greatest disadvantage is with error-correcting, when compared to key-to-tape or terminal input, which requires merely backspacing and retyping. Tab cards, furthermore, are limited in general to 80 columns, whereas key-to-tape and input terminals are without this limitation. The advantage of tab card input is its independence of the computer. Marked sensed cards and optical character readers have not lived up to the hopes of potential users, mostly because of imposed limitations and relatively high costs. The ultimate answer may be direct voice input, but as of now we are wedded to the traditional tab card.

With the advent of magnetic tape, the flexibility and speeds of computer input/output or I/O were increased measurably. Magnetic tape is normally a polyterephthalate ribbon, ½ inch wide, 1.5 mils thick, and 2400 feet long, coated on one side with magnetic iron oxide. The density of information is commonly 800 bits per linear inch (800 BPI), although 1600 BPI tapes are being used increasingly, with seven or nine tracks. The advantage of magnetic tape is its high storage capacity, about 20 million characters or 5 million words, at the lowest cost (about $30); its disadvantage is that it must be read serially. In general, tab card input is transferred to magnetic tape before being introduced into the computer for processing.

From the viewpoint of the computer and of the user of the computer, the job to be done is delineated when the inputs, manipulations on the stored data, and outputs have been described in appropriate detail for the specific type of computer. Communications with the computer is by means of programming language for which there is software (assembler or compiler) that converts it into machine language. The programming language is not a means by which a user "talks with" a computer, but essentially a sequence of detailed and explicit sets of directions, instructions, or commands for accomplishing predetermined objectives. In high-level languages—e.g., FORTRAN, COBOL, PL/1, etc. —these directions, commands, or instructions are very much like English. COBOL and PL/1 are generally the most useful for chemical information systems.

Programming is the backbone of computer applications. But relatively few literature chemists are familiar with any programming language, and of these only a small fraction is expert. Consequently, most literature chemists must use computers through programmers who have no knowledge of chemical information systems and how the systems are used by and useful to chemists.

Whether programming is through a programmer or by our own effort, we need to define the problem, to analyze the problem in terms of operations, to program the problem as a series of operations or manipulations, to delineate the output, and to document the problem so we can solve it again. The objective is really to program the problem so that the computer is instructed to take in, store, work over, store, and give out data in the way we need the data. As literature chemists, we can hardly hope to achieve the computer designer's and maintenance engineer's knowledge of how the electronic circuitry works, the systems analyst's skill in designing software; and the programmer's experience in writing programs. Yet the more a literature chemist knows about the electronic circuitry, the software, and programming the better the computerized information system will evolve from his knowledge of the data going in and the results issuing from the computer by means of the logical outline of the programs.

LOGIC AND THE COMPUTER

Because of the phenomenal speed of the computer, software designers and programmers may not always consider the various aspects of logic that affect the efficiency of computer operations. There are many ways of operating on or manipulating data in computers that produce the same results, but not with the same efficiency. Of particular importance to tasks in chemical information systems is the sort/merge operation that arranges information in alphabetical or numerical order.

Sort/merges in the core or with peripherals, such as magnetic tape, employ a variety of techniques, such as exchange sort, radix sort, sieving, etc. Tape sort/merges require a minimum of three tapes. Whether or not a proposed logic is the most economical over-all must be evaluated against software and programming costs. The important point is that there are many ways of solving problems in terms of computer logic. Thus, it is important to consider the logic of programming instructions if the computer is to be used at its maximum efficiency.

DESIGNING A CHEMICAL INFORMATION SYSTEM FOR COMPUTER PROCESSING

Thinking, evaluating, and decision-making are the first steps in setting up a computerized chemical information system. We need to know and to understand the information problems and needs of our user group associated with a set of documents. We need to determine the proper content and scope of the input that can be derived from the set of documents. Proper content and scope means the kind and amount of information that satisfies the needs of the users with relevance and convenience and at a cost that can be justified. The design needs to be such that the users are not inundated with massive outputs that contain the one or two pieces of pertinent information. Nor should it be such that the user is challenged to "ask the right question." Thinking, evaluating, and decision-making are the important consequences of computerizing a chemical information system.

Computerizing chemical information systems does not require that we have to become programmers or that we have to understand how the electronic circuitry of a computer works, but, as pointed out before, it surely helps. It does require, however, that we define and organize the information and logic that will become the basis for the computer operations. Because a computer can select and permute keyworks in a title easily and quickly is a poor reason for basing a computerized information system on this method.

Data Type	Content of Data
1	Subject S1
1	Subject S2
1	Subject SN (Nth subject)
2	Title of Document ....
2	....
3	Abstract ....
3	....
3	....
4	Author A1
4	Author A2
5	Location of Author(s)
6	Reference
7	Discipline of Science Classification
8	R&D Classification

Figure 2. An input design

Record No.	Data Type	Content of Data
X	1	Subject S1
	2	Title of Document
	3	Abstract
	3	Abstract continued
	4	Author 1 + Author 2
	5	Location
	6	Reference
	7	Discipline
	8	R&D
X	1	Subject S2
	2–8	Same as above
X	1	Subject SN
	2–8	Same as above

Figure 3. Subject storage format

Nor should it be assumed that what a computer does easily and quickly is necessarily economical. Keyword indexing requires computer look-up storage for elimination of words and, in more sophisticated systems, for thesaurus control of words; in addition, the permutation logic and operations are relatively high in their consumption of core storage. In short, keyword indexing is a fairly expensive computer operation and yields a product of questionable retrieval merit—i.e., in terms of GIGO, garbage in garbage out.

There is only one way to avoid GIGO. That is to select input records intellectually, eliminating and minimizing garbage before it can pollute the system, and to design retrieval concepts from the viewpoint of the needs of users. We are then ready to set up a computerized system.

Four related aspects need to be considered in setting up a computerized system:

1. Input capability
2. Data storage capability
3. Processing capability (software and programming)
4. Output capability

With reference to the journal literature system discussed earlier, these four aspects are related initially by the input design. Figure 2 illustrates one input design with which all four aspects may be flowcharted. Many options are open to us. Let us assume the simplest input and storage options. Each selected document is analyzed for the types of data indicated in Figure 2 and keypunched, one tab card per line with the data type in a fixed column—e.g., column 3—and the content (subjects, title, abstract, etc.) beginning in a fixed column—e.g., column 5—and restricted to a maximum of 60 columns. So that we can control the data storage capability, we need to establish a reasonable record length for the maximum input per document, such as 1500 characters (equivalent to 20 tab cards, 75 characters or columns per card; the average input, however, is less than 10 cards).

The simplest storage option from tab cards to magnetic tape is illustrated for subjects in Figure 3. Thus, separate tapes could be dedicated to subjects, authors, author locations, references, etc., for updating, sorting, and outputting in predetermined schedules. This option, however, is only one of many.

CONCLUSIONS

The computerized information system is now a familiar concept. Although the computer is an elegant tool, it is one which is easily misused intellectually and economically. When used with intelligence, skill, and knowledge, it is a tool that can add new dimensions to chemical information systems as well as result in appreciable cost savings over traditional systems. But thinking, evaluating, and decision-making and the bringing of vision and creativity to the chemical documentation aspects are the important factors in using computers efficiently and productively for chemical information systems.

BIBLIOGRAPHY

(1) Brooks, F. P., Jr., and K. E. Iverson, "Automatic Data Processing," 466 pp., Wiley, New York, N. Y., 1969.

(2) Cole, R. W., "Introduction to Computing," 336 pp., McGraw-Hill, New York, N. Y., 1969.

(3) Golden, J. T., and R. M. Leichus, "IBM 360 Programming and Computing," 342 pp., Prentice-Hall, Englewood Cliffs, N. J., 1967.

(4) Jordain, P. B., and M. Breslau, "Condensed Computer Encyclopedia," 605 pp., McGraw-Hill, New York, N. Y., 1969.

(5) Katzan, H., Jr., "APL Programming and Computer Techniques," 329 pp., Van Nostrand Reinhold Co., Princeton, N. J., 1970.

(6) Williams, W. F., "Principles of Automated Information Retrieval," 475 pp., The Business Press, 1968.