By JAMES W. PERRY

Chemists and Librarians

Dr. Perry is professor of chemistry, Massachusetts Institute of Technology.

Also das sind die Aufgaben der wissenschaftlichen Forschung, die Feststellung von Tatsachen durch Beobachtung und Experiment, das Sammeln und Ordnen dieser Tatsachen und ihre zusammenfassende Beschreibung.—P. Jordan, *Die Naturwissenschaften.*

THE spectacular advances made by chemistry during the past quarter century have tended to obscure a crisis which has been developing in that important field. The nature of this crisis is best understood by considering how the science of chemistry advances and develops.

New experimental findings are, of course, the *sine qua non* of progress in chemistry. Yet it is most unusual for any one experiment to extend the range of knowledge by more than an infinitesimal amount. Standing alone, any one experimental fact is virtually insignificant. Each experimental observation, each new fact, does not achieve a full measure of value until it has been correlated with other observations and facts.

It is not always immediately apparent, however, which observations are actually related, and the development of comprehensive theory may require the review of a wide range of facts scattered both with respect to time and to the place in which they are recorded. Furthermore, industrial development of any one branch of chemistry may suddenly and unpredictably require a large amount of information concerning some other branch of chemistry previously regarded as quite unrelated. Thus, for example, information concerning the molecular structure of kerosene hydrocarbons became important in the development of synthetic detergents. The colloidal chemistry of various soaps dissolved in gasoline was important during the war in developing flame throwers. Successful application of chemical knowledge often requires the collection and correlation of facts previously widely scattered in the record of chemistry.

The record of chemistry, however, continues to expand year after year as a result of the enormous amounts of time and money devoted to chemical research. It is becoming increasingly clear that the volume of recorded chemical information is approaching a point at which the value of the record in its present form must be impaired by its very extensiveness. The approaching crisis is causing some chemists to feel considerable concern for the future development of chemistry, both pure and applied. As a consequence, attention is being directed to the possibility of applying new methods and new devices to chemical information problems.

The possibilities inherent in new devices are best understood by first directing attention to those mechanical devices which—like the straight-edge and compass of classical geometry—have been the conventional tools of information work. Due to the conventional character of such tools it may seem somewhat strange to regard them as mechanical devices. Nevertheless, from an objective point of view, conventional indexes involve use of a mechanical device, namely,

bound sheets of paper on which lists of words are arranged. Pigeon holes and other shelving devices constitute a mechanical means for isolating items into grouped arrangement for classification purposes. Card files of conventional type, considered as mechanical devices, consist of a large number of separate pieces of sheeted cellulose capable of being arranged with convenience in one order without possibility of being rapidly and conveniently rearranged in some other order.

An enormous amount of thought and effort has been devoted to working out procedures, systems and rules for achieving the maximum of accomplishment with simple conventional devices and, indeed, much has been accomplished with these simple tools. Existing methods of indexing and classifying, based on the use of simple conventional devices, have been used for such a long time, however, that many persons regard such methods as involving rules which have an absolute validity and which can be extended to other mechanical devices. It is not generally realized that the introduction of newer devices into information work opens up possibilities which we can exploit fully only if we are willing to devote considerable time and effort to re-examining the basic problems involved and to working out new solutions.

The situation might be compared to the relationship between development of new weapons and the evolution of new tactics in warfare. Thus, the introduction of air weapons confronted military commanders with the necessity of revising ground tactics. So it is with the introduction of radically new devices in the scientific information field. Using the old, well-known indexing and classifying methods in conjunction with powerful new modern devices for handling information would prove as shortsighted as it would have been not to revise 18th century tactics for maneuvering foot soldiers in our modern air age.

Approach of Chemist

From the chemist's point of view, the starting point for reevaluation of the chemical information problem is the fact that the day is long since past when any one chemist could hope to read all the available chemical information of possible eventual interest to him in his work. It would be even more impossible to retain in his memory the enormous mass of useful detail involved. Human limitations have forced chemists to transfer to an even greater degree the important function of memory to bound volumes stored on the shelves of libraries and to collections of more or less private reports.

The function of memory, however, implies much more than the mere act of storing, however neatly or efficiently it may be accomplished. A collection of numerous volumes and reports devoid of pathways leading to desired information is of little value. In fact, the value of such a collection depends on the effectiveness of the means available for arriving at desired information.

For many years chemists have trod at least three different paths in searching out published chemical information. The indexes to abstract periodicals have constituted one of these paths. Summarizing compendia such as Beilstein and Gmelin have been another. A third path has been provided by texts devoted to various aspects of chemistry. Each of these means for locating chemical information has proved valuable and yet at the same time each has been found to have its own peculiar limitations.

The usefulness of abstract periodicals is limited by the time of effort required to consult the indexes, look up the original abstracts, and take notes. Not too long ago, when the number of bound annual volumes

of the abstract periodicals was relatively small, the amount of time required to consult the record of chemistry in this fashion was considerably less than it is at present. Looking into the not too distant future, it is possible to anticipate that the labor involved in consulting indexes and abstract journals will become excessive.

The great compendia of chemistry have proved themselves invaluable. They have the disadvantage, however, that the classification scheme followed is often not in harmony with the point of view of the person seeking information. Thus, if one wished to obtain from Beilstein a list of all organic compounds known to have poisonous properties, it would be necessary to check the entries for each individual compound. The reason for this is the fact that classification in Beilstein—and in Gmelin—is based on molecular composition and structure rather than on properties. Such classification is useful to a person interested in locating compounds having certain properties, e.g. color, poisonousness, only to the extent that he knows or correctly surmises that certain groups of compounds have the property of interest. The situation is aggravated by the fact that the grouping of compounds in compendia is often different from that required for locating certain information. The reason is the fact that all possible groupings of compounds based on molecular composition and structure are not presented in compendia because of practical limitations arising from the cost of printing on bound sheets. The indexes of abstract periodicals provide even less, with respect to grouping of compounds and related information, than is the case with the compendia. Here again the rigidity of ordering of printed sheets in bound volumes and the cost of printing impose limitations best characterized as mechanical in nature.

Books as a means for locating information often prove unsatisfactory for similar as well as other reasons. Books are usually written from a rather specific point of view to cover a specific range of subject matter. It very often occurs that no book has been written which covers a subject of interest to a chemist. Books, moreover, rarely treat information pertinent to a given subject in an exhaustive fashion and even when, in the exceptional case, coverage is complete at the time of publication, obsolescence with respect to newly published information is rapid.

The time and effort required to consult the record of chemistry has become so great that many chemists prefer to copy information once it has been located onto cards or into notebooks, and thus gradually build up small collections of information whose principal merit is convenience, *i.e.*, the speed and ease with which items of information can be found again when desired. It should be noted in passing that this copying of information often wastes the time of highly skilled persons capable of more creative effort. It is scarcely surprising, therefore, that chemists in general regard searching the literature and the attendant taking of notes as distasteful drudgery. This attitude can and often does lead to overlooking important information with inevitable wastefulness in conducting both research and development work.

As mentioned above, many chemists maintain files of information gleaned from the record of chemistry. Perhaps a non-chemist might think it a simple matter to maintain and use such files. Actually this is not the case, due to the fact that papers and reports in the field of chemistry usually deal with a number of closely interrelated yet nevertheless distinct matters, such as details of reactions and syntheses, properties and uses of substances, theoretical questions, etc. It is often surprisingly time-

consuming to search even a relatively small file of chemical information for the purpose of providing the answer to some scientific or technical question. As a consequence, the use of punched cards, in particular the edge-notched, hand-sorted type, has become widespread among chemists. By suitably punching these cards the subject matter written on them may be characterized simultaneously by a number of criteria. Sorting of the cards may be based on any one of the criteria indicated by the punching or by any combination of such criteria. This principle of multiple designation of characterizing criteria with resultant flexibility in selection of items of interest, has repeatedly demonstrated its usefulness in the chemical information field. In fact, punched cards have proved so useful in managing small files of information that the possibility of applying them on a much broader scale could scarcely fail to attract attention. The American Chemical Society has had a committee actively studying the matter since January 1946.

Problems

Preliminary investigation has revealed not only encouraging possibilities of accomplishment but also a series of problems, some of them novel in nature.

One of the more obvious problems is that presented by structural formulas, particularly those of organic compounds. These formulas—occupying a position with respect to chemistry somewhat akin to that of wiring diagrams in the field of electronics—might be described as the picture language of molecular architecture. Conceivably a machine might be devised to handle this form of picture writing directly. For the foreseeable future, however, such a machine appears likely to remain in the realm of theoretical possibilities. Since practical devices, available at present or in development, can handle words much more effectively than hieroglyphics, it might be thought that the names of compounds rather than their structural formulas could be used as the basis of machine sorting. It so happens, however, that present practice in assigning words to designate chemical structures is not on a satisfactory basis for the purpose under consideration. The rules of chemical nomenclature in their present state of development contain too many exceptions and only too frequently result in ambiguities and uncertainties. If the efficiency attained by machines in manipulating numbers is to be extended to molecular structures, then a system for completely and unambiguously designating such structures in terms of convenient symbolism must be developed. The English chemist, Dyson, assisted by collaborators both in this country and abroad, has developed a system for completely representing any given organic molecular structure by a linear array of symbols consisting of letters, numerals and punctuation marks. Plans are now being developed to demonstrate experimentally that automatic equipment can carry out sorting and searching operations based on use of the Dyson system. It is anticipated that these experiments will reveal many as yet unsuspected advantages to be gained by employment of modern mechanical devices in the chemical information field.

All chemical information cannot be expressed, however, in terms of structural formulas, whose function, as already noted, is limited to expressing the architecture of molecules. Chemists also make use to a certain extent of mathematical symbols and concepts definable in mathematical terminology. In additon, they use a large number of words whose meaning may not be as sharply defined as one might wish. Evidently, semantic problems will be encountered in developing a system for me-

chanically searching chemical information.

Not quite so evident is another problem arising from the previously mentioned need to re-examine previously developed indexing and classifying methods. As already noted, the most valuable feature of punched cards and similar devices is their ability to record a plurality of characteristics in such a way that searching operations may be effected not only on the basis of any one of the indicated characteristics, but also any combination thereof. Full exploitation of this valuable feature is fundamental to efficient use of punched cards and related devices in information work.

Consideration of a very simple example points to an important and as yet incompletely solved problem. Suppose we are concerned, not with chemical information, but with newspaper stories. Among these there would be many items involving a dog, a man, and the act of biting. If our punched-card coding is to achieve its maximum usefulness it must be able to distinguish between "Man bites dog" and "Dog bites man." If, moreover, a state of insanity should be involved in some of the news accounts, we would doubtless wish to be able to distinguish between "Mad dog bites man," "Mad man bites dog," "Man bites mad dog," "Mad man bites mad dog," etc. This simple example may suffice to show that, for maximum effectiveness, coding for punched cards and similar devices must indicate relationships between concepts and not be limited to recording the mere simultaneity of index entries. Stated somewhat differently, sentences rather than words must form the basis of coding. Although in theory it might be possible to construct a machine that would be able to scan sentences written in English and respond to them in a desired fashion, it seems obvious that the problem of the machine designer will be simpler (and the final machine much less expensive) if the sentences serving as the basis of coding are written in a standardized fashion. If this were done the machine would be required to respond to a standardized pattern of relationships having none of the irregularities, idioms, etc. which characterize English and other human languages. It should be noted that the machine language could be made quite complex as long as it were kept free of irregularities. Development of a "machine basic grammar" is a problem now being investigated.

Another aspect of the problem of using machines to search files of chemical information involves the relationships between concepts. It seems likely that certain common features of concepts might well be indicated by suitable coding just as involvement of the Greek root meaning "to write" is indicated by the spelling of certain English words, such as phonograph, biography, telegraph, geography, ideograph. Relationships of a semantic rather than a philological nature will almost certainly play an important role in machine searching of information files. How best to weld together semantic relationships and machine techniques is a problem that will require much careful study.

Summary

Ready availability of information is a most important matter to the chemical profession. Continuing expansion of research activity will not accomplish what it should if improvements in means for making information available do not keep step. The application of new tools to information problems requires their reevaluation. Development of new methods for indexing, classifying, etc., are essential if the promising possibilities of modern machines are to be fully realized.