gain in speed and efficiency from mechanization and the use of machines is lost if the material which goes into the machine has not been skillfully and dependably selected to present the facts accurately and fully.

The literature chemist will gain, in my opinion, by recognizing clearly the opportunity which is open to him, bringing to the work all the intelligence, ingenuity and insight of which he is capable, doing his work with meticulous detail when meticulous detail is needed but never losing sight of the over-all canvas, and coming together to discuss problems, objectives, ways and means as we are doing here. This has been a great help and will I am sure continue to bring status and stature to those in this work.

As for compensation, the fastest way in which to get higher pay will be to advance to an executive. From that position, one can have a voice in determining salaries. When your turn comes, don't forget.

The literature chemist has interesting work, demanding work, important work. The problems awaiting solution are formidable; he should stand up and be counted.

Information Retrieval and the College Chemistry Curriculum*

By ELBERT G. SMITH
Mills College, Oakland 13, California
Received June 6, 1962

In recent years there have been geologic upheavals in chemical literature. The hills are being thrust up into mountains and these will soon become mountain ranges. The horse and buggy methods that were once good enough for getting around in gentler terrain are becoming less useful and soon will become impossible. There seems to be considerable recognition of this in chemical industry where something is being done about it by developing new methods for retrieving chemical information. As yet there seems to be little recognition of, or interest in, this problem in most of the colleges and universities in this country. Less than half of them even offer a course in chemical literature and only a very small proportion of the publications in the chemical information retrieval field is coming from academic institutions. I should like to discuss some of the reasons for this situation and what might be done about it, but first it seems desirable to review for my academic colleagues some current developments in chemical information retrieval that are taking place in industry.

As a chemical company's file of compounds grows, it becomes increasingly evident that the difficulties of nomenclature and the catch-all nature of molecular formula indexes make indexes based on these principles increasingly helpless, or expansive, or both, in dealing with the company's requirements for information. New kinds of indexes are being tried, such as coordinate indexes or permuted indexes. Machines are being used to do information retrieval jobs that were formerly impossible and even unimaginable. In at least one company a chemical notation system is being used for indexing company compounds and this index has completely displaced both nomenclature and molecular formula indexes. I shall illustrate this general movement within industry with examples from a small area of chemical information retrieval with which I happen to be most familiar.

One pressing problem in the pharmaceutical industry and in a growing number of others is that of generic chemical structure retrieval—locating all those compounds in the company's file that have the same functional groups, or ring structures, or chain structures, or logical combinations of these features. Indexes based on nomenclature are utterly powerless to cope with this problem and even inverted molecular formula indexes, once thought to be a solution, aren't really of much use for most structural features and in any event bog down when an index grows to any considerable size. Files of 10,000 to 50,000 compounds are not at all unusual in these industrial concerns and at least one company forsees a file of 100,000 compounds.

The most common approach in industry to this problem of generic structure searching has been the fragmentation of structural formulas into bits and pieces which can be assigned code numbers which in turn can be manipulated with punch card machines of varying complexity to produce lists of the compounds that contain the particular combinations of structural units required in a given problem. Most of these codes trace their descent, one way or another, from the pioneering work of Donald Frear, Karl Heumann and others who developed the CBCC code[1] more than 15 years ago, and from the work of Fred Whaley[2] who simultaneously simplified these practices and developed advanced ideas of logical manipulation. Starker and Cordero[3] recently have published a typical example of one of these fragmentation codes and Geer and Howard[4] have listed other examples that have been proposed and sometimes used.

Another approach to this problem was pioneered at Monsanto Chemical Co. by W. H. Waldo,[5] who has devised a way of putting edited structural formulas directly into a computer and of searching these stored structures with a suitable computer program to print out the structures of those compounds meeting the search requirements. This approach is being developed in at least two other companies.

A third approach has been pioneered by Howard Bonnett[6] at G. D. Searle & Co. and was described by him a year ago before the Division of Chemical Literature at the St. Louis meeting. In this approach structural formulas are translated into Wiswesser notations which then can be arranged, by the simpler kinds of punched card machines,

* Presented before the Divisions of Chemical Literature and Chemical Eduction, American Chemical Society 141st National Meeting, Washington, D. C., March 21, 1962

into a variety of printed lists in which the notations are arranged in different ways. These notation lists, bound in book form, make it possible for a chemist who knows the notation to make successful generic searches at his own desk, without further use of a machine, by visually scanning the notations in particular sections of particular lists.

These approaches to this problem of generic structure searching are practical and pragmatic responses to pressing needs within industry but it is not difficult to see a role that college and university chemistry departments might well play in this development. If these systems increase in number, and many people think they will, trained people are going to be needed to operate them. This will be particularly true if chemical notation systems once gain some acceptance, as they very well may because there is now evidence[6] that they are the easiest and quickest, and therefore the cheapest way of doing a number of jobs, including generic structure retrieval, provided one knows the notation system.

Quite aside from the training of people at this rather technical level, it also is apparent that the devising of new information systems and the improvement of old ones will require people with a degree of training and insight that is surely equal to that required of the most creative bench chemists. In addition, the practices resulting from these approaches to generic structure searching often raise broader questions than those they were designed to answer—questions that often are beyond the immediate practical concerns of information chemists in industry. For example, are the highly varied fragmentation codes now being devised and used in various chemical companies susceptible to integration into a common code that might be useful to everyone? What about the economics of these various approaches? Is it really necessary to go to a machine each time a generic structure search is required? Can or should a final word ever be reached in devising a chemical notation system? How can chemists be persuaded that learning a notation system would simplify some of their information retrieval problems? Add to these problems the necessity of developing adequate teaching materials to explain these new systems to the beginner and it's apparent that this one small area of chemical information retrieval alone is bristling with problems of a long-range or educational nature that might well be better attacked in an academic chemistry department.

There are many other problems in chemical information retrieval besides this one example. How can an abstractor choose words for an abstract that will satisfy both the chemist who needs specific information and the indexer who wants to help the chemist find the abstract? Indeed, how can one arrive at the best practices for selecting words or terms for indexing? And once you have the terms, what kind of index is best made with them? Coordinate indexes, for example, though quite familiar to industrial chemists, seem to be less available to academic chemists. This is partly because on the industrial scene there are useful published coordinate indexes dealing with chemical patents and also because some companies have developed coordinate indexes of their own for intramural use. How and for what materials could coordinate indexes be made that would be useful to academic chemists?

Problems of this sort could and should be investigated in the sort of intellectual environment that colleges and universities traditionally offer, where one is free to work on long-range problems of one's own choice and where immediate practical applications may not be the compelling and limiting factor they often are in industry. In the academic environment there is often more opportunity for stimulating contact with people in other disciplines. This is particularly important in the present state of information retrieval practices where ideas from a variety of disciplines, from writing and semantics to symbolic logic, may well be as important as the chemical knowledge brought to bear on an information retrieval problem. The chemical knowledge, however, must be the basic material on which these other disciplines must operate.

Then where's the rub? Why have so few people in academic chemistry departments seemed to have sensed the excitement of a fascinating, challenging and important area of study? There seem to me a number of reasons for this, some the fault of the chemist, some the fault of the information specialist and some because of inadequate communication between the two, especially on the academic scene. Indeed, how can the academic chemist become familiar with these matters I've discussed? The *Journal of Chemical Documentation* is going to help mightily here. As I've tried to show in the reference list at the end of this article, the January 1962 issue of this new journal offers one of the best entries that I know to current work in this field. In the past, however, it just seems that academic chemists have had little contact with the other journals in which information specialists publish and even if they do occasionally come to the chemist's attention, he may well be repelled and confused by some of the things he finds there. In some of these publications the information specialists or the editors or both seem to have little notion that some sort of experimental evidence ought to back up the claims made for the information retrieval systems described in their pages. A quite expensive study has had to be made in the past year by the National Research Council for the National Science Foundation on just what is actually being accomplished by current chemical information retrieval systems, in large part because in some of the published literature it is quite impossible to separate the actual accomplishments from the pipe-dreams.

It may be partly because of this unfortunate situation that some academic chemists seem to feel that information retrieval isn't quite academically respectable—that the methods of the information specialist have little in common with, say, thermodynamics, or the elucidation of structure, or other traditional topics, and have too much in common with the trades of the warehouseman, the plumber or the embalmer, who aren't taught their crafts in a college or university. I choose these three crafts advisedly since information specialists are concerned with setting up the pumps and pipes to get carefully pickled information in and out of some sort of storage device. It may even be that sometimes there is an element of craftiness in the devious means some chemical information people must use to get answers to questions for which their information retrieval systems weren't designed to cope. Nevertheless, academic opposition to information retrieval as a mere craft doesn't make much sense, coming from chemists, since the information man's craft-like

means are employed, not as the plumber's or embalmer's, but just as the chemist uses the crafts of the balance or buret—to produce results of *intellectual* concern.

I think that academic chemists are more justified in objecting to another kind of thing in which some information specialists indulge. During the past year I've had a rather unusual opportunity to listen to what information people have to say about what they do and to talk with other chemists who have been listening to them too. I've a strong feeling, that seems to be shared by other chemists, and I suspect by anyone trained in the outlook of the physical sciences, that often, when I listen to the information specialist, I'm sinking in some sort of amorphous goo of words that as often as not leave me baffled as to what, if anything, the information specialist has said. It seems to me that at least one reason for this is that some of these people are using some of the words of physical science without understanding what they mean to the physical scientist. The word that causes me the most trouble is the word "theory." To the physical scientist this word stands for a rather high order abstraction, usually a *set* of hypotheses, as in the kinetic-molecular theory or Dalton's atomic theory, that relates a group of generalizations called laws. The laws in turn grow from the experimental verification of more tentative generalizations which are usually single hypotheses. The hypothesis is a more or less informed guess as to a possible relationship that may exist between experimentally verified data of one sort or another. When some information people begin talking about information "theory" and it all too soon becomes evident that they have not yet gone far enough in the exploration of their field to discover any laws, it is no wonder that people from the physical sciences get that feeling of sinking in an amorphous goo. In other words, when the information specialist says "theory," he had better be saying "hypothesis" if he hopes to keep the physical scientist from feeling that information people are suffering from delusions of grandeur. The point I'm trying to make is that there is rather a lot of difference between the statements, "This paper is of *theoretical* interest," and "This paper is of *hypothetical* interest."

If this analysis is correct, it implies that information retrieval is not yet a science since it does not yet consist of laws linked by theories. It has practices but not laws. It is still in the stage of accumulating data about itself. It is in the stage that chemistry was nearly 200 years ago. We knew how to weigh accurately enough to collect a lot of information about what weight of one substance reacts exactly with what weight of another, but we hadn't yet found the laws of combining weights or conservation of mass. Generalizations such as these cannot come out of thin air, however, at least not in a science, but arise out of a body of verified facts. Information retrieval is now more or less in the stage of accumulating verifiable experiences. In the area of chemical notation systems, for example, some of us are in the stage of making alphabetized notations lists of considerable size. Howard Bonnett has made printed lists of 20,000 Wiswesser notations and 30,000 more are in the works. I have a file of 50,000 Wiswesser notations on punched cards which I have shown how to search generically for chemical structures on a simple punched card sorter (7). I doubt that anyone could have predicted the remarkable ease with which such lists can be scanned visually to find desired structures until such lists were available for study. Another phenomenon that both Dr. Bonnett and I have noted is that incorrect notations have a strong tendency to accumulate at the "interfaces" between groups of related compounds. They stick out like sore thumbs and make such errors easy to detect and correct. Again, such an observation could not be known until these notation lists had actually been made. I've no doubt at all that there are other properties of notation lists waiting to be discovered and this is another area that might well be explored by information chemists in academic institutions.

Information retrieval may not yet be a scientific discipline but this only accentuates the need to bring to it the discipline of the scientist. In a young and growing field this does not always happen. For example some information people seem prone to accept uncritically the conclusions of others without any critical examination of the evidence on which such conclusions are based.

I know of at least one large enterprise where great effort is being expended to do an information retrieval job in a particularly difficult way because the information man in charge is relying in part on just such a published conclusion. It supports his own prejudices and presuppositions, but it is not justified by the experimental procedure used, the method of carrying it out or the interpretation of the data obtained.

On the other side of the fence all is not healthy either on this account. Some chemists who begin working in information retrieval seem unable to bring to bear their previous scientific training which presumably has taught them to look for verifiable evidence before drawing conclusions. For example, too many people have told me, "We looked into chemical notation systems and decided against using them," when subsequent conversation more often than not revealed that they, or one of their hired hands, may have skimmed through a notation manual, once over lightly. If you were to ask them, "How many thousand compounds of what kinds, or from what sources, did you encode into what notation; what did you do with the notations after this and what was there about the results you found unworkable?" they would be greatly taken aback. They haven't dreamed of doing this sort of thing. Yet this is the kind of experimental evidence that should surely be necessary to give some meaning to such a decision. I don't want to dwell on these human lapses. After all in a young and growing field of study we must expect growing pains. Nevertheless we must recognize that some of this unscientific hanky-panky rubs off on information retrieval as a whole and a college curriculum committee may be somewhat justified in looking down its collective nose at proposals to teach something about information retrieval in a college chemistry department.

Nevertheless, in spite of these difficulties, it seems to me that most aspects of information retrieval do fit into the college and university frame of reference. A chemical information specialist may well be involved in writing, abstracting or symbol manipulation and this sort of thing is already being taught, in one way or another, in many colleges and universities right now. The new approach

to mathematics through the concept of sets is familiarizing students with concepts of logic and symbol manipulation that surely will find more and more application in information retrieval techniques. The college in which I teach also offers a sophomore course in report writing which includes considerable experience with the writing of abstracts. Two of our younger English department faculty are offering a graduate seminar in semantics and surely these ideas should play some role in the writing of clear abstracts, in developing thesauri and other indexing and searching aids. Incidentally, the reason that semantics is being offered by these younger men is that some of the elder statesmen of that department look down on semantics because there has been a crackpot fringe about it.

This is a good example of the kind of academic opposition we can expect to the teaching of chemical information retrieval within a chemistry department unless somehow, some of the more unfortunate practices of some information people can be abated—the misuse of words that have definite meanings to the scientist, the publication of pipe-dreams with no foundation of experimental evidence, the uncritical acceptance of the other fellow's conclusions and the idea that snap judgments can somehow substitute for verifiable experiment.

Finally, what can happen within the colleges and universities themselves which may help this somewhat awkward child to grow and mature? I suspect this will follow the path of any other new discipline. One person in a chemistry department will become interested, as a matter of intellectual excitement, in these new ideas about the storage, manipulation and retrieval of chemical information. He will have to read and study and learn pretty much on his own. Some of his colleagues will feel that he isn't really doing research at all. Ideas will occur to him that he'll want to try out. Successful experiments will lead to publication and if this soundly done, he may begin to get some recognition, perhaps even research grants and eventually perhaps he may attract a few graduate students who have a nose for a new and exciting area of knowledge and who, at the same time, want to get a foot in a ground-floor door of a new discipline. This may well lead to the discussion of information retrieval techniques in an expansion of an existing chemical literature course, and perhaps later to a separate course as the field grows. These students probably will find it advisable to support this work with courses in other departments in writing, abstracting, symbolic logic and the like.

Such a curriculum will emerge only if it has the patient nurture of administrative officers with imagination to sense new directions of intellectual effort and if there is forthcoming the necessary financial support from industry and the foundations. In such an atmosphere a chemical information specialist should be relatively free from the compulsions in most industrial jobs to hurry up and produce a practical information retrieval system that works a little better than the one the company has been stuck with for the last umpteen years—only to be stuck with the new system for umpteen more because of the cost of the change. In the academic setting the chemical information specialist should be considerably freer to explore, discard when necessary and explore again. This may well help speed the time when someday, somewhere, someone will see the first tiny crystal of an honest-to-goodness generalization form in that amorphous goo; a scientific law will finally have been discovered and information retrieval will be on its way to becoming a science.

LITERATURE CITED

(1) National Research Council, Chemical-Biological Coordination Center, "A Method of Coding Chemicals for Correlation and Classification," Washington, D. C., 1950.
(2) Whaley, F. R., *Amer. Doc.*, **12**, 101–107 (1961).
(3) Starker, L. N., and Cordero, J. A., *J. Chem. Doc.*, **2**, 12–15 (1962).
(4) Geer, H. A., and Howard, C. C., *J. Chem. Doc.*, **2**, 51–53 (1962).
(5) Waldo, W. H., *J. Chem. Doc.*, **2**, 1–2 (1962).
(6) Bonnett, H. T., and Calhoun, D. W., *J. Chem. Doc.*, **2**, 2–6 (1962).
(7) Smith, E. G., *Science*, **131**, 142–146 (1960).