*Pure & Appl. Chem.*, Vol. 49, pp.1897-1900. Pergamon Press, 1977. Printed in Great Britain.

# TECHNIQUES FOR THE RETRIEVAL OF CHEMICAL INFORMATION; DREAMS VS. NIGHTMARES

F. A. Tate

Chemical Abstracts Service, Columbus, Ohio, U.S.A.

Abstract - Computerised data bases are immature and understanding of how they may best be used is low. For the future we must look towards more effective mapping of data base content onto user requirements and the development of automatic procedures for assisting the user to get maximum response from information systems. This will require co-operation between data base producers to co-ordinate practice and between them and systems operators to create integrated systems for access to required information.

To put my comments in the proper perspective for the audience, let me explain that I was a chemist, but I am not one now. I am not a computer man. I am not an economist. I do not claim to know what an information scientist is, but I do claim that I am not one. I am an information partisan and an optimist.

I am going to take an overview rather than comment on each of the papers individually and in dealing with the future I will emphasise some of the factors that will lead to change. Please recognise that I, like most of you, heard a great many things that were not said. I also believe that many things were said which were not heard or at least were not interpreted as the speaker had intended by most of the audience. The point is that an expert often fails to be understood when discussing his speciality with a non-expert audience. For instance, consider the discussions of structure-handling. The speakers' candor and the ability to use the Queen's English is beyond question, yet I believe that their responses to questions were not understood by the questioners. The jargon of these responses was interpretable only by the very few specialists in computer-based systems that handle chemical structural diagrams who are present today. Certainly the level of communication in these cases was less than the speakers wanted.

There were several discussions of the evaluation of systems. Comparisons of systems without reference to the functional intent and the working environment of each system leaves a considerable doubt as to the value of the comparison. Most of the systems that were compared were created for different purposes. They live in different frameworks and they serve different audiences. The systems which were compared do overlap in use, but they do not overlap in a very substantial way. For instance, some of the information systems which were compared provide service to the public; some are operated for a captive audience; and some serve both types of audiences, but not to the same extent. In all cases, economic comparisons are misleading without carefully identifying what is being compared.

At the present time data bases are very immature. They have not existed long enough to accumulate sufficiently extensive data bases to permit anyone to conduct a comprehensive search. Few go back more than ten years. Thus, a search for a fact reported in the primary literature prior to the mid-nineteen sixties must usually depend wholly upon printed tools.

Thus data bases have been created to produce printed information-accessing services and the content of such data bases reflects the organisation and redundancy of printed text and corresponding indexes. Data base content is also too limited in depth, because it is included only when such content is readily accessed through search of the corresponding printed service. The time has come to include content which may be useful only through automated search and which may, therefore, often not be included in the printed counterpart to the data base. And as Arthur Elias pointed out, there is very little inter-data-harmony at the present time. And user problems in dealing with overlapping content of data bases are numerous and complex.

Information centres have revolutionised the use of computer-readable information services. After rapid increases in the number of data-base installations during each of the first half dozen years that computer-readable information services were available, the number of installations started to decrease. The reason for the reduction has been that getting a large data base ready for automated search requires a substantive investment and unless these

make-ready costs can be spread over a large number of searches, the cost per search is too high to be acceptable. The result is that organisations are shifting from running their own searches to using centres which serve many organisations, and centres which serve a limited audience are dropping by the wayside. Networks have also contributed to the reduction in the number of information centres by providing ready access to data bases installed at remote locations. This makes geographic separation of data bases acceptable; this contrasts with the emphasis on the concentration of resources which we associate with traditional library collections. Such separation could in the future become a severe burden to information users if the natural tendency for information centres (i.e., or network nodes) to develop widely variant search personalities is not curbed. At the present time, the user of on-line data bases located in two or more different network nodes pays a high use tax in the form of the burden of having to be familiar with several search protocols. This not only reduces effective use of these tools, it reduces the efficiency of centre operation.

Information users are seldom trained in the use of computer-based accessing services, therefore, they must use intermediaries to search on their behalf. There are several reasons for this circumstance. The most obvious reason which causes information users to go through intermediaries is the unfamiliar and complex interface to the search system itself. The interface to the system has none of the comfortable familiarity of library use. The problems of unfamiliarity could be partially alleviated by giving these systems proper emphasis in present curriculums of higher education. Such inattention results in part from the newness of computer-based information systems, for they are largely unfamiliar to those who are well established in the hierarchy of the scientific and technical communities. Aside from the unfamiliarity of most information users with the potential of automated search systems, the complexities of formulating searches is imposing, for automated search is not conducted in natural language. It is negotiated in terms of parts of words and fragmented grammar. A search profile has lost nearly all of the normal syntax of natural language. Instead of being a precise statement of facts needed, a profile is more like a fishing net, and it operates with all the precision of such a net.

There is a common misconception that chemical information services are only for chemists. I would estimate that over 60% of those who depend on CAS services to meet their information needs are not chemists. Moreover, few information users depend on just one accessing service to meet their information requirements. Thus a great many searches can only be satisfactorily carried out by using several data bases, perhaps each being accessed through a different on-line supplier. Within large organisations intermediaries who specialise in specific data bases may be practical, but in smaller organisations such specialisation is impractical. The result is that the complexities of training intermediaries to search two or more data bases, each accessed through a different on-line centre employing its own specialised search protocols, make it near impossible for many information users to get effective access to needed information.

Search intermediaries are usually considerably less expert in the search subject than the information user himself. This then requires the user to explain to the intermediary what information is needed. The user expresses in his own words the detailed characteristics of his information problem. In contrast, if the information user were to conduct his search personally through printed publications in library facilities, he could express his information requirement with words used by the original authors of pertinent papers. Thus, use of current systems often lacks the spontaneity of library search and it can be inconvenient and time consuming. As computer-search systems become ever more important by providing substantive improvement in the user's access to information, the impediment of complex interfaces to search systems, unless they become vastly simplified by automation, will stand as a barrier to scientists' use of the information storehouse.

On-line search multiplies the problems of educating those who conduct the searches. Because each terminal with its operator constitutes a mini-centre, the number of those who actually conduct searches has been dramatically increased by the inception of on-line searching. Also, as each terminal provides access to several, perhaps many, different data bases, there is a likely chance that the searcher will be inexperienced in the search of all the data bases he uses. Thus, the qualifications of the intermediary at many of the terminals are often not as good as they have been in many off-line search centres. At the present time the number of on-line service suppliers is relatively small and each provides access to several data bases, usually with the similar search protocols applicable for all data bases which a given on-line supplier provides. Education of intermediaries is manageable in today's working environment, because up to now each of the on-line service suppliers has managed to train its search clients adequately. But as the number of available data bases becomes larger, as the individual data bases grow larger and more complex, and as the number of information users who depend on data base access increases to the worldwide scientific and technical community, it will not be possible to operate effectively with the existing practice of depending upon search intermediaries.

I would now like to look at some of the intermediary functions which on-line information services will have to supply automatically in the future. The system itself must offer the

inexpert searcher the chance to check unfamiliar details of search formulation and operation. It must prompt such users automatically. On the other hand, the experienced searcher must have the option of skipping over unneeded formalities. Profile encoding, that is the conversion of a subject question into computer-oriented search code such as Ernest Hyde described in discussing substructure search, should be automatically accomplished with the system prompting the user in avoiding potential ambiguities and to advise the searcher of potential problems such as the consequences of imprecise profile definition. The search encoding support must also provide automatic incorporation of search strategies to reduce the cost of operation and assure system responsiveness. And in shifting search from one data base to another, the searcher must be automatically alerted to differences in preferred terminology, data element definition, format variations, and any other variations in editorial prerogatives.

The elements of the search profile must be automatically edited by the interface system. Such edits must deal with: system conventions of spelling and abbreviations; preferred terminology; equivalence of synonyms, acronyms, and symbolic representations; identification of potential ambiguities; and reliable identification of unaccepted variants to these details including likely misspellings in search terms. The system must negotiate with the searcher through graphic display or in natural language as the searcher chooses.

In effect, the searcher must utilize almost all of the editing capabilities and the numerous dictionary files which the producer utilizes in building the data base. As it would be impossibly expensive to implement these kinds of processor tools as part of each on-line service centre, the searcher must be provided with access to dictionary files which already exist in on-line form within data-base-producing operations.

More use must be made of encoding-decoding techniques in search files to reduce size, to increase search efficiency and reliability, and to improve use economics. Encoding-decoding techniques have already proven effective in handling chemical substance representations in information files and similar techniques must be applied to bibliographic and subject vocabulary data. These techniques will provide the basis for bridging among data bases and will provide the basis for efficient handling of overlap among data bases by both processors and data base searchers.

On-line service suppliers must become more conscious of the audience which it services. Every centre must map audience's use of each of the data bases to which it provides access to as to eliminate unuseful content and to seek out the additional content which will assure retention of its audience. Also the present high cost of updating on-line search file places an unreasonable economic burden on users of these services. This comes about partially from the immaturity of present data bases and partly because of the distinct differences in search file structures from one on-line supplier to another. Without economic, reliable updating, search files will not be kept current.

Processors must coordinate their practices in document identification so that centres and users can easily and reliably identify overlapping coverage of documents. Also, until now, the data base supplier has created the definition of the data bases which it produces. This mode of operation severely restricts data base versatility. By contrast, the Library of Medicine chose to acquire file content from a number of data base producers and integrated this input into the TOXLINE data base. This will become a common means of building data bases of the future. Data base producers must coordinate their practices so that each centre is able to pick and choose among content from overlapping data bases in building a data base which meets the special needs of that centre's audience. This approach to synthesis of data bases implies the need for coordination among data base producers at the data element level.

Hyde referred to the beginnings of substructure search based on fragments. The move to computer-based topological search resulted from the problems of updating manual fragment search files. All of today's topological search systems utilize screens which are in fact no more than fragments such as were used in the first-generation, non-automated substructure search systems. Please note that I consider linear notations and systematic nomenclature to be logically equivalent to each other and to connection table representation of topology. The experience of Internationale Dokumentationsgesellschaft für Chemie (IDC) and BASIC which is a subsidiary of Ciba-Geigy, Hoffman-LaRoche, and Sandoz of Switzerland shows that nearly 100% screen out is possible without any loss of current answers. As Hyde pointed out, once a topological file is established, screens can be changed at will. As for the future of substructure search, I believe that the emphasis will return to screen-only (i.e. fragment only) search. And I must note that screen-only search is readily compatible with current text search services.

Let me touch briefly on the handling of numerical data which have been the focus of many of the presentations. Not much such data appears in present information-accessing systems or even in the primary literature. Most data of this type remain only in the files of laboratory results within the investigator's parent organisation. Occasionally experimental data are placed in a depository associated with the primary journal in which the corresponding

paper is published. Because of the lack of coordination among such depositories and the complete invisibility of data retained in the author files in his home laboratory, backup data is inaccessible or available. Therefore, experimental data have little chance of being re-used until they are: automatically recorded in computer-readable form as part of the measuring operation; carried forward as part of the automated record in the data reduction, analysis, and rationalization processes; archived in such a way as to make them visible and accessible in computer-readable form to those within the scientific and technical community who are interested. The American Chemical Society is starting to produce its primary journals through an automated system. This provides a first step in building an uninterrupted link back from the information-accessing service directly into corresponding primary documents. The system needs to be extended to permit the link between journal papers and supporting laboratory data.

Information users cannot cope with a wide range of automated information services unless attention is given by those who produce the services to the need to use data bases effectively in combination. I do not suggest that there should be only one accessing system. There must be many kinds of data bases, many routes of access to information and many specialised services. And information users must be given the capability of using the techniques developed for manipulating data bases which are distributed to the public in handling their proprietary information.

In discussing what "must" and what "will" occur in information processing, I may have left some doubts about what the future holds for information systems. Let me conclude by stating that I have no doubt that all changes which I referenced will occur. Great advances can be expected in the next five years, but some may take a decade or more. And just so I am on the record, I do not believe that printed copies of CHEMICAL ABSTRACTS will die out in the next couple of decades.