*Pure & Appl. Chem.*, Vol. 49, pp. 1889 - 1896. Pergamon Press, 1977. Printed in Great Britain.

# A GENERAL REVIEW OF CHEMICAL & OTHER INFORMATION SYSTEMS OF RELEVANCE TO USERS OF CHEMICAL INFORMATION

A. W.Elias

Biosciences Information Service, 2100 Arch Street, Philadelphia, Pennsylvania 19103, U.S.A.

Abstract - The material presented takes a problem oriented overview of its' subject, focussing on the communications activities of chemistry and those of related subject areas. Emphasis is placed on the functions that are accomplished by a variety of participants in the communications pathways and the changes in the functions or their location that may take place in automated systems.

A series of analogies are drawn in order to project the quandries that are presented to researchers who attempt to obtain interaction among and between chemical and related information retrieval systems. These analogies project a number of potential solutions, suggest which participants are best suited to provide the solutions and some of the practical problems that these participants will have to meet.

Some possible programs and mechanisms to meet these problems are suggested.

The title of this presentation is "A General Review of Chemical and Other Information Systems of Relevance to Users of Chemical Information". This is clearly no title for an abbreviated presentation. It is the title of a book, a doctoral dissertation, a monograph - - almost anything but what it has to be in these circumstances - a distillation of the essentials and a record for the future. Even without the papers presented at this conference, the volume of material is enormous, and the timing of this paper, does not allow me to produce a pattern of chemically specific problems so that you can correlate them with the solutions you have heard here.

If my title has led you to expect an exquisite analysis of information systems, past and present, annotated and analyzed for every scintilla of information potential - you will be disappointed. The approach that I employ does not separate out systems, services and publications, based on their chemical information constituents, but deals with the overall problems of relating chemical and other information sources. From this we can consider strategies and problems and gain some insight for practical solutions.

Let's examine the major functions that go into the creation and maintenance of most present-day "information systems", by document processors for these activities are basic to future systems interaction.

FIGURE 1

DOCUMENT PROCESSOR FUNCTIONS

COMPACTION

EXTRACTION

ADDITION

INDEXING

STANDARDIZATION

There is a function that "compacts" information. The degree of such compaction varies from the "informative" to the "indicative" abstract with a host of special forms.

Another function may be said to be "extractive" where certain information is made more prominent according to the special goals of a particular service.

In order to pick the locks, that is enter the data base, requires that group 3 tacticians decide not only on the data bases and their sequence, but on the amount of brute force to be applied to each. Many users can make a choice and if properly educated, can match different command languages to the specific data element and file organisations of selected files.

Once this selection is made, further choices exist for our burglar in determining the specific tools from his kit. These tools include the various indexing approaches, classifications, thesauri and the like. Not surprisingly there are preferences and the SDC report examined these as well.

When the searchers were asked their preference for various methods of accessing data bases, the responses were limited to vocabulary structures of a few defined types. It is regrettable that this limitation was made for it would be interesting to evaluate other approaches specific to chemistry (e.g. formulas, structures, etc.) The report describes the following types of vocabulary structures:

CONTROLLED VOCABULARIES

FREE LANGUAGE VOCABULARIES

COMBINATIONS

Controlled vocabulary terms are assigned by indexers in reference to a thesaurus or authority listing, while free language vocabularies are based on indexer selection of terms from document text. While more current, the lack of controls develops a greater variety of terms. There are of course certain "Vocabularies" that are totally "derivative" in that no human (indexer) term assignment is made.

FIGURE 2

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 1. I HAVE THE MOST SUCCESS WITH SEARCHES PERFORMED ON DATA BASES WITH . . . | 23.2% | 48.5% | 8.0% |

FIGURE 3

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 2. I HAVE LEARNED MOST QUICKLY ABOUT THE COVERAGE AND SCOPE OF DATA BASES WITH . . . | 40.1% | 23.2% | 6.9% |

FIGURE 4

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 3. I AM MOST EFFICIENT (TIMEWISE) WHEN PERFORMING SEARCHES ON DATA BASES WITH . . . | 39.9% | 30.4% | 7.4% |

An "additive" function also exists. Here the document processor installs alternate forms of information (e.g. chemical codes) and supplements or enriches the information.

Both the "additive" and the "extractive" functions serve as input to still another function, that of indexing where the goal is to provide for later access. Indexing is central to the success of any interface between chemical and related systems as we shall see.

A crucial function which has long been recognized, but erratically put into practice, is that of "standardization" so that commonly useful points are made identical.

With this cursory review of systems functions, we can climb to a higher perch for an overview. This peak is called Mt. Interaction.

Systems interactions is not a new problem. Kent and Geer (1) in addressing problems of "Searching Chemical Information Mechanically" discussed "reference" searching, indicating the use of agreed upon terminology to connect a new system with earlier bodies of information, and the problems involved when agreed upon terminology does not exist. They concluded: "Information-retrieval systems engineered to serve the various levels of literature searching may be drastically different from one another. Appraisal of the potential information retrieval requirements of the foreseeable future may be a critical consideration in engineering a new system". The Kent - Geer paper was written in 1956, 20 years ago, and it is easy to document similar plaintive cries for interaction.

Mellon in his book, "Chemical Publications" (2), addressed the problem in his chapter on "Making Searches in the Chemical Literature". This chapter is laden with solutions and good advice on how to surmount the incompatability of chemical literature. It progresses through a veritable labyrinth of inconsistencies involved with indexes, chemical nomenclature and numbering schemes. One must sympathize with Mellon and suspect his frustration when he includes the following footnote:

"The Student should not forget that, even though certain facts cannot be found in published works, he ordinarily has two other possible methods of obtaining the desired information: he may inquire of the individual who knows what is desired, or he may resort to experiments in the effort to determine the facts for himself (!)" - - I do not know if this footnote is in Mellon's first edition, (I consulted the fourth), but he was surely, if sadly, right in 1965. We are probably in the same position in 1976.

This leads us to examine the tools that exist or can be projected in order to accomplish interaction among and between chemical and "related" systems. We can consider several approaches. I call the first:

The Sequential Lockpicker

This tool is that most currently used. I have called it the "sequential lockpicker" because it brings to mind the mode of operations employed by a burglar who enters a building which contains money and finds that it is concealed behind a series of locked doors. He must determine which doors to open and how to pick each of a variety of locks.

In a recent survey of the impact of on-line retrieval services, Wanger, Cuadra and Fishburn of the System Development Corporation (3) addressed the selection of data bases for a search. They found that 80% of the searchers reported access to on-line searching services involving more than one bibliographic data base. These searchers reported on the tactics employed in selecting appropriate data bases for a given search.

There were five tactics indicated for "most searches"

| | | |
|---|---|---|
| 1. | Use one data base for one search | 48.2% |
| 2. | Use a second data base when tactic 1 is unsuccessful | 12.1% |
| 3. | Select different data bases that are relevant and try the search on each one | 43.3% |
| 4. | When number of "hits" on one data base is not as great as expected, try the search on another file | 15.8% |
| 5. | Regardless of results from tactic 1, search on another to confirm results or to gamble on finding something relevant | 8.8% |

We can see from these tactics the pertinence of the "sequential" part of the title of this tool. For the "lockpicker" part of the analogy, we will focus on the problems employing tactic three.

FIGURE 5

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 4. WHEN PERFORMING SEARCHES IN SUBJECT AREAS IN WHICH I AM MOST COMFORTABLE OR KNOWLEDGEABLE, I PREFER DATA BASES WITH . . . | 16.9% | 48.7% | 13.3% |

FIGURE 6

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 5. WHEN PERFORMING SEARCHES IN SUBJECT AREAS IN WHICH I AM NOT PARTICULARLY COMFORTABLE OR KNOWLEDGEABLE, I PREFER DATA BASES WITH . . . | 34.0% | 39.3% | 7.5% |

FIGURE 7

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 6. WHEN PERFORMING SEARCHES ON DATA BASES WITH WHICH I HAVE HAD PRIOR EXPERIENCE (E.G., THROUGH CODING FOR BATCH-SYSTEM SEARCHES), I PREFER DATA BASES WITH . . . | 15.6% | 33.3% | 7.2% |

FIGURE 8

SUBJECT VOCABULARY PREFERENCES

| | CONTROLLED TERMS ONLY | CONTROLLED PLUS FREE TERMS | FREE TERMS ONLY |
|---|---|---|---|
| 7. I PREPARE MORE FOR SEARCHES ON DATA BASES WITH . . . | 32.9% | 20.3% | 19.9% |

The contexts in which the questions are asked affect the judgements given and are worth review. When the premium is for success in searching (question 1) then the favoured preference is for a combination by almost 2 to 1. 3. When the emphasis is on ease of learning, the controlled approach is favoured by about the same proportion. 4. Efficiency is thought attainable in about equal proportion in controlled and combination vocabularies, 5, but subject background or experience in a given data base leads to a preference for the

combination vocabulary. 6. In an unfamiliar subject area, there is little preference for one approach over the other. The "tools" are therefore not necessarily selected on the basis of the job to be done, but on the basis of the competence of the craftsman.

7. Direct prior experience in off-line methods favours the combination approach, while 8. more preparation seems to be needed for the controlled vocabularies. The use of free terminology (without indexer intervention) is the least favoured approach no matter what context is employed. This may provide some insights into the potential uses of free text from primary sources as a system input.

It should be noted however, that this survey was conducted mainly among users of the MEDLINE system, and all of their training favoured the use of controlled vocabularies (supplemented by a free language capability). Chemical searchers might give very different responses.

The next interactive tool that our burglar can employ may be termed the "inside job". Here we assume that our burglar is given the key to one room in the building and upon entering it, finds the key to another room and so eventually can obtain access to all of the rooms in the building including the one with the money in it. Once in any given room, he can search all of its contents (there may be a few dollars lying around) and he is assured of at least one route to another room. This is the method currently used in order to connect chemical files with related ones. In order to interact then, the key retrieved in one room, must be a key that will search another. Some knowledge is also required as to the sequence of the rooms to be searched and of course as one acquires keys, he is increasingly able to improve his likelihood of success.

The approach to interaction that is involved in the creation of compatible files and the use of standards is the "inside job" approach, and many data base suppliers devote much time and effort to placing one or more keys in their data bases to facilitate this type of burglary. There are other strategies possible.

Let us consider for example the use of an accomplice. If our criminal has a confederate who knows a lot about the construction of the rooms and the locks, the process changes considerably.

The accomplice strategy assumes minimal alterations of the files. In a software sense, an accomplice could be simulated through so-called "table look-up" programmes to open the doors.

In another form, we have the "master key strategy". In this case, the master key is outside the building and independent of the rooms or data bases. Such a "master key" approach is used in plans for interaction through the development of "authority files". These would also act without altering the data in any given room or file, but would group all sorts of keys (synonyms, codes) and promote file interaction.

Finally, we could consider the Dr. Moriarty approach. In this case, a master criminal genius is located nowhere near the scene of the crime. His work produces many solutions to our locked door puzzles and all of them are direct, easy and cost effective.

Employing file mapping, graph theory, statistical association, linguistic analysis and other exotic tools, it maybe possible to provide maximal interaction among and between chemical and related data bases.

This slide attempts to tabulate the analogies for present and future developments of interactive tools. For each criminal "Modus Operandi" or "MO", I have indicated the tool(s) for providing system interaction among and between data bases. The next column attempts to indicate the requirements for its use or development.

For the "Sequential Lockpicker" "MO", the corresponding systems approach is simple. We need do nothing, because this is basically the situation that now exists. Even if this is permitted, however, there are some requirements to make even this situation endurable. We will require an increased level of data base education in order to have any reasonable chance to succeed. This is coming about in programmes supported by data base suppliers and by the purveyors of systems used to work with these files on-line. This type of educational effort costs money and the question of who pays has not been resolved.

If we adopt the "inside job" "MO", we will need the installation of common indexing keys in the chemical and related files. To accomplish this, we have additional requirements. We will require someone, to assume responsibility for installing common keys. Next, there will have to be agreement as to which common keys will be installed. Once this occurs, the nature of the keys must be addressed. There will also be a requirement for good, workable, well thought out standards that can be employed for this purpose. Education in data bases will still be required in this approach so that the selection of information can be made with complete understanding. Of course, there will be costs to do this.

FIGURE 9

DEVELOPMENT OF INTERACTIVE TOOLS

| CRIMINAL "MO" | INTERACTIVE "MO" | REQUIREMENTS |
|---|---|---|
| "SEQUENTIAL LOCKPICKER" | NO CHANGE IN CURRENT PRACTICE | DETAILED EDUCATION<br>$ INVESTMENT |
| "INSIDE JOB" | COMMON INDEXING KEYS | ASSUME RESPONSIBILITY<br>OBTAIN AGREEMENT(S)<br>SELECT KEYS<br>DEVELOP STANDARDS<br>$ INVESTMENT |
| "ACCOMPLICE" | INDEX KEY CONVERSION(S) | ASSUME RESPONSIBILITY<br>OBTAIN AGREEMENT(S)<br>OBTAIN COOPERATION<br>$ INVESTMENT |
| "MASTER KEY" | DEVELOP "AUTHORITY FILES" | PERFORM RESEARCH<br>ACQUIRE DATA<br>$ INVESTMENT |
| "DR. MORIARTY" | FILE MAPPING<br>STATISTICAL ASSOCIATION<br>LINGUISTIC ANALYSIS | PERFORM RESEARCH<br>$ INVESTMENT |

The "accomplice" "MO" requires a capability for the ready conversion of Index Keys. In this approach, there will also be a need for assumption of responsibility. Here, the suppliers seem most likely candidates. The cross-disciplinary elements involved in working with related files will suggest resource sharing to achieve file compatibility. And, such sharing implies supplier co-operation.

For the "Master Key" Approach, system interaction will probably employ so called Authority files. This term cannot be subjected to an absolute definition at this time. Many "Authority" compilations are now being made available. Upon closer scrutiny, a number of them are index key convertors described in the Accomplice MO, but true authority file development is an area of important value. In a time context, an authority file could fill gaps between the current and prior data. They could operate as independent entities, with great potential impact on interaction in the bibliographic data bases they support.

The requirements for these tools include voluminous data. In order to serve their purpose, their construction must be most ingenious so that they can interact with dissimilar collections. We are in the very early stages of forging such tools and research and development groups will and must investigate the concept if they are to help.

For the Dr. Moriarty "MO", I have indicated just a few techniques under systems interaction, file mapping, statistical association,and linguistic analysis are potential methods. The posture of the "MO" is data independent. These inquiries relate to the structure of both the information and the nature of the inquiry process. Its results can be applied at the point of the burglary or as easily be inserted in the design of the building, the rooms, the locks and even the keys. This type of research costs money.

The role of networks to create interaction regardless of the type, seems to me to be minimal. Its potential to force interaction however may be a major factor. As the networks and the users they serve grow in numbers, the amount of pressure for interaction will increase in growing proportions.

## CONCLUSIONS

Although I did not disembowel any particular chemical or related system, I did look carefully into the characteristics of a great many of both types. As you have heard, I found serious and challenging problems, but I also found ingenious, resourceful and dedicated minds who are solving the problems and meeting the challenges.

If we take an inventory of the different requirements that unfolded as a result of the overview for interactive tool development, we can see that it includes detailed education, assumption of responsibility obtaining agreements, the selection of candidate keys for interaction, development of standards for those keys, co-operation among participants in the processes, research, data acquisition, and dollar investments.

FIGURE 10

REQUIREMENTS FOR INTERACTIVE TOOLS

DETAILED EDUCATION

ASSUMPTION OF RESPONSIBILITY

OBTAINING AGREEMENTS

KEY SELECTION

STANDARD DEVELOPMENTS

COOPERATION

RESEARCH

DATA ACQUISITION

DOLLAR INVESTMENTS

Of course, these requirements have built-in inter-relationships and are not independent variables, but it is convenient to summarize them as if they were independent, to estimate how they have developed in the past, seem to be proceeding currently and what may lie in the future.

Starting with detailed education, this requirement is so needed to save the sequential lockpicker situation in which we are today, that is receiving much attention. In the past, when most of the files were available only in printed form, education for their use was heavily concentrated in the training that particular scientific groups obtained in the course of their education. The amounts of this training and the emphasis that it received was dependent on the importance that each discipline attaches to the information. Chemistry, by its very nature requires much more of this than many other disciplines, and yet even the record in chemistry has been spotty as indicated by Mellon's sad quotation. Unfamiliar and/or new approaches have a difficult time in penetrating the consciousness of end users.

Currently we see a changed level of emphasis. The growth of machine readable information tools in both off-line and on-line environments, and their availability for use side by side, has been a major contributor to the changes. Beginning with the intermediaries, those who work with the tools as their main daily occupation, the demand for education has been growing. And, as the end-users become more and more involved, we can anticipate demands from this group as well.

Responses to these needs have come from both document processors, and data processors who are offering training programmes for detailed data base and system education.

For the future, the educational approaches will evolve to match the development of the interactive tools. As the other strategies for interaction are developed, other forms of education will be needed and of course the content of that education will change.

Assumption of responsibility as a requirement for the development of interactive potential is less clear. The participants who would be expected to take major roles in assuming responsibility lie in many areas. Their interest, and the amount of power each possesses to assume responsibility vary widely.

Unifying organisations such as IUPAC itself can and have assumed active roles. The suppliers, that is the document processors and data processors, can assume responsibility of course, but they will take a narrower view to meet the interests of their services first.

In the absence of a master plan that would organise all of the resources available, the future course is difficult to predict. For some time to come, self-interest will affect the assumption of these responsibilities and it is a matter of whose self interest is involved that will govern the forms and the programmes themselves.

Once responsibility is assumed, it may involve only a single information service, in which case it can be implemented. Obtaining agreement then has been a matter for unilateral decisions and this has been the pattern of the past. Currently, the technical rewards that can be achieved require that relationships between information services grow stronger and more intimate. In addition, to the technical adjustment that must be contemplated, there is also a strong sociological component involved. There are gradations relating to the particular levels of agreement and the amounts of flexibility, that any given group can allow. Still there is progress. I can cite our own experience at BIOSIS with programmes involving Chemical Abstracts and Engineering Index in evidence and there are many other examples.

For the future, I believe that we can look to steady growth in obtaining such agreements.

The selection of candidate keys for interaction is a major requirement and as might be anticipated, its very importance has limited progress and programmes. It is here that it is difficult to arbitrarily make a selection that will serve the needs of chemical and related systems. From a chemical point of view, emphasis on a chemical point of interaction would seem most desirable and the Chemical Registry number programme is a leading candidate. Still, the related systems will have different priorities. In order to achieve such selection will require sacrifices (mentioned in my discussion of assumption of responsibility and obtaining agreements). As we will review in just a moment, the economic factors will also have great importance. Possibly the future for such selection will require a co-ordinating body to aid in selection or, as seems more practical, a focus on accomplishing selection of simple things. This approach was taken by BIOSIS, CAS and Engineering Index in our work on document description keys.

Once the keys have been selected, it will be an absolute requirement that the form in which they are represented be standardized. If this is not done, we will have only compounded confusion at just the point where it is to be eliminated. Here we again see the possibility that it will be easier to standardize on simpler and more universal elements. Co-operation among the participants in this variety of activities has been developing over many years aided by the work of interested groups. While many worthwhile studies have been performed, the localized interests of the groups have acted to limit overall programming.

For the future, opportunities must be found that will bring these groups and the interest they represent into some more organized approaches to the work. While bilateral agreements can do much in providing for specific levels of interaction, a voice that speaks powerfully for multi-national, multi-disciplinary and multi-functional interests has yet to appear.

Research in the technological and theoretic areas that would support new approaches to interaction has been shifting its emphasis from the technological to the theoretical in recent years. The change in emphasis from "how" to do the compacting, extracting, and indexing, steps to an understanding of "why" the processes are needed, and the development of a basic theory of information science is evident in the increasing amount of mathematical notation used in papers in the field. As data bases become more available for experimental uses, the researchers in the field can demonstrate the applicability of theory to practice in more and more realistic experiments. This work will provide the long term input to the Dr. Moriarty "MO" and its encouragement and constant scrutiny of its findings for application to the problems of interaction is of great importance.

Data acquisition for the provision of authority files has been made more practical with the use of so-called "active-storage" by many participants in the promotion of interaction and the tools that it requires. Problems still exist relative to the incorporation of older data that is vital to the comprehensive files that will be needed, yet which is in a form that will require special processing. Methods for the accomplishment of this kind of work at reasonable costs have not received the attention that their future employment will require.

Finally, the economics of the range of work required are always with us. Projects and programmes receive their support from such a great variety of sources that it is difficult to tabulate them. While the interest of governments, professional societies, data base suppliers, and end users are concerned, the funding (except in rare cases) has been sporadic and the results parallel the funding. If an overall planning activity could be promoted, at the very least there could be some organized projections of the funding needed and the possibility of obtaining the funds could be approached in terms of a plan as opposed to specific projects or programmes.

In closing, the requirements of modern societies for effective interaction among and between information files and systems can be met through technical, societal, educational and economic factors so that communication patterns of chemical and related information systems can be magnified and improved. Peter Drucker possibly put it most succinctly by saying: "What people want most is a little mobility, a little freedom from the constraints of a traditional society, and a little information that links them to the world."

REFERENCES

1. Kent, Allen and Harriet Geer, Searching Chemical Information Mechanically, in Searching the Chemical Literature. Advances in Chemistry Series. 30:270 - 281 (1961) R. F. Gould, Ed.

2. Mellon, M.G. Chemical Publication: Their Nature and Use. 4th Edition. McGraw-Hill Book Co., 1965

3. Wanger, J., C.A. Cuadra and M. Fishburn. Impact of On-Line Retrieval Services System Development Corporation, Santa Monica, CA 1976.