in the abstracts in CAS' *Chemical-Biological Activities* computer-readable file. CNIC expects to make this integrated text-structure search system available for general use in France.

Because CAS is the only single comprehensive chemical information system, its continued financial viability is important to all nations who wish to build their future information activities on this 72-year-old record of chemistry and chemical engineering. The financial support from the organizations in the agreement countries assists greatly in preserving this viability. The responsibility to assure the continuation of the service is thus shared among organizations in five nations rather than resting only on the shoulders of the ACS and the United States. There is one difference in the basic financing of the agreement country operations and those of the ACS. The difference may become more significant in years to come. The organizations in each of the agreement countries have received financial assistance from their national governments. There is no such support, at present, for the ACS information operations from the U.S. government, although in the 1965–75 period CAS did receive some $25 million for R&D.

Along with the obvious financial advantages and the joint R&D efforts which have been described, this five-nation international cooperation provides the support of respected leaders and organizations in the four countries and helps to build understanding of mutual interests, philosophies, principles, and policies. The organizations bring local expertise into marketing efforts in their nation's culture and environment which would not be possible otherwise. They also provide a base for user education and training for all types of local scientists and their supporting staffs.

Undoubtedly, one of the most valuable benefits from these international agreements is the opportunity to interact with large organized groups of users of CAS publications and services. This is something which has been very difficult for CAS in the past. CAS subscription files contain addresses of many intermediaries, librarians, book dealers, etc., but few actual bench scientist users. The agreement organizations are supplying very necessary interactions with these end users. IDC has been outstanding in these efforts. They represent a large group of industrial scientists and operate an efficient and growing patent information service. Their constructive suggestions have included a wide range of interests: all the way from the simple addition of volume numbers to abstract numbers in each printed CA abstract, a very popular improvement, to complex improvements in the CAS patent coverage and indexing policies. Similarly, the UKCIS has assisted materially in the improvement of new *CA Selects* profiles and in the documentation, instructions, and plans for improved on-line services. The French and Japanese agreements are relatively new, but equally important user assistance is anticipated from these two major groups of international scientists.

CAS began by sharing the responsibility for abstract production with a large group of international volunteer abstractors. While their activity has decreased in importance, the international sharing has grown via four existing bilateral national agreements which provide marketing, user interactions, R&D assistance, consulting advice, and financial support. This international sharing is expected to grow in the years to come and to be of even more importance in the total chemical information system.

Present and Future Prospects for Structural Searching of the Journal and Patent Literature

JOHN A. SILK

Imperial Chemical Industries Limited, Plant Protection Division, Jealott's Hill Research Station, Bracknell, Berkshire, England

Received May 14, 1979

Available systems for structural searching of organic compounds in the journal and patent literature are briefly reviewed. From a survey of the various methods of substructure search it is concluded that an algorithmic notation able to deal with both specific and generic descriptions of structure could be an extremely valuable development, since it would occupy a key position intermediate between canonical connection tables for individual compounds and degenerate fragment coding for groups of compounds.

The question, "How can we best attain comprehensive substructure searching of the journal and patent literature for low molecular (nonpolymeric) compounds?", is an important practical one for industrial information services in the pharmaceutical, agrochemical, and general organic business areas. The development of on-line searching has had a major impact on the retrieval of text-based information, and this includes chemicals which can be identified adequately by names or registry numbers. It has, however, only served to emphasize the limitations of available systems for searching generically for classes of compounds or those containing specified structural units.

To set the problem in perspective we must examine its component parts, namely, the available data bases and the systems for searching them, and, moreover, we need to make a distinction between past and future. Whatever new systems and data bases might be developed over the next few years, there is little likelihood of improvement in the keys by which the older literature can be searched. Consequently, for the past we need to identify the best of what is available, and to consider how to utilize it most effectively. For the future we need to examine the current state of affairs and try to decide what developments would be most desirable.

PRESENT-DAY SOURCES

Table I lists the major compilations which cover the journal and patent literature in a manner providing some degree of retrospective search. Foremost is *Chemical Abstracts*, which covers both journals and patents and provides an excellent

0095-2338/79/1619-0195$01.00/0

Table I. Coverage of Major Chemical Data Bases.

Source	Journals	Patents	Specific	Generic
Chemical Abstracts	✓	✓	✓	✓
Derwent CPI	—	✓	—	✓
Index Chemicus*	✓	—	✓	✓
IFI/Plenum*	—	✓	—	✓
IDC	✓	✓	✓	✓
PDR	—	✓	—	✓
BASIC	✓	✓	✓	✓

retrieval capability for individual compounds, but only a limited one for classes of compounds. Next in importance is Derwent Publications Central Patents Index which deals only with patents and provides only a generic search capability.

Thirdly, there is ISI's *Index Chemicus Registry System*, which through the use of Wiswesser Line Notation provides both a generic and a specific search capability. Its coverage is, however, limited to first mentions of new compounds or improved preparations in a core of about 110 journals.

Fourthly, there is the IFI/Plenum data base of U.S. patents. Only in recent years has this provided a generic search facility based on fragments, and its limitations to U.S. patents is a serious one for European and Japanese companies.

These are the only four primary indexing sources for the organic chemical literature, three being American and one European. The journal literature is covered by two, and the patent literature by three. There are, however, three other European systems which rely on a reprocessing of material produced originally by Chemical Abstracts Service and Derwent.

The most important of these is produced by the German organization IDC. This covers both journals and patents, and its GREMAS code provides a very detailed and precise substructure search capability. Next is the Pharmadokumentationring, usually referred to as the Ring, which comprises a group of European pharmaceutical companies who until recently shared the cost of re-indexing Derwent material into their own fragment code. Thirdly, there is the Swiss consortium BASIC, which, like IDC, purchases *Chemical Abstracts* tapes for in-house reprocessing. Operational costs have not been published.

In Britain three years ago the Association of Information Officers in the Pharmaceutical Industry (AIOPI) set up a study group to examine ways of improving patent searching. Through the cooperation of IDC and a member of the Ring, a series of searches were carried out to compare retrieval on different systems, and details were presented at the 1978 Derwent conference;[1] the studies have been extended a little further since then. From them three conclusions may be noted.

(1) While the early Derwent fragment code had a number of deficiences, which resulted in very large numbers of false drops, the current version is, on balance, comparable in performance with Ringcode. This was shown not only by our tests but it is also the view of the members of the Ring, and it lay behind their decision to save the growing costs of recoding Derwent material and to stop at the end of 1977.

(2) The GREMAS code of IDC possesses a far more precise search capability than Derwent or Ringcode. Its ability to deal with specified positions of substitution on a ring or chain, and to distinguish between substituents on different ring systems within a complex molecule, is particularly valuable in some well-trodden fields of patent activity.

(3) Although all three systems displayed a high degree of recall—and this despite varying degrees of relevance—none was superior consistently in providing all the known answers.

The AIOPI group recommended, therefore, that consideration be given to putting the Ringcoded backlog of Farmdoc and Agdoc on-line beside Derwent's own multipunch coding.

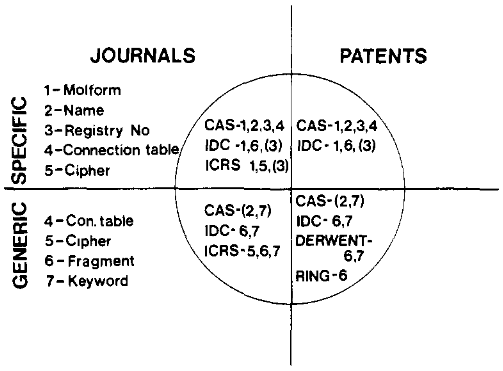


Figure 1. Search keys of chemical data bases.

As Ringcode is already familiar to subscribers of Ringdoc, and more recently of the Chemical Reactions Documentation Service, this would entail less effort in learning an additional coding system than any other. The approach is now being studied by a subgroup of the Derwent Chemical Coding Committee. The main problem to be overcome is that of time overflows in on-line searching, because of the number of heavily posted terms for commonly occurring features, but it is hoped that it may be available by next year on the U.K. Infoline service.

It is also worth remembering that IDC is now part of the German National Information Center for Chemistry and as such it offers a public search service of its combined patent and journal file for a fee of 1500 DM. On occasions this can be invaluable. Another advantage is that the data base covers starting materials and intermediates as well as products.

For retrospective substructure search of the journal literature we consider in ICI that the tapes of the *Index Chemicus* Registry System provide a valuable data base. This file of about two million compounds goes back to 1965, and since all the compounds have been ciphered into WLN, they are searched by our CROSSBOW system. There is, of course, the alternative of using the Chemical Substructure Index based on permuted ciphers produced by ISI.

FUTURE DEVELOPMENTS

When we turn to examine possible future developments, Table I can be represented in the form shown in Figure 1. The circle denotes the total domain of low-molecular organic compounds. It is divided vertically into two halves which denote the journal literature and the patent literature, and horizontally into two halves which represent the potential for searching it specifically for individual compounds and generically for classes of compound.

Within each segment are shown the organizations whose data bases cover it and the means by which they may be searched: molform, name, registry number, cipher, connection table, fragment, and keyword (*keyword* is used here in the broad sense of any alphabetic or numeric descriptor, be it a single term, a subject index entry, or a manual code or patent classification).

Only one organization is fully represented in all four segments, namely, IDC. *Chemical Abstracts* provides mainly a specific search capability, but it is only partially represented in the lower (generic) part by its subject and molform indexes, which are useful in some situations but not in others. ICRS is limited by its chosen coverage, while for patents Derwent and the Ring provide only a generic search capability.

If we regard our objective as being one comprehensive system, it is worthwhile to examine some of the alternatives which this suggests.

Firstly, we could all support IDC. This solution is (in principle) available immediately. If IDC had as many subscribers as Beilstein, perhaps we could all afford it!

Secondly, ISI or another organization could be asked to cipher all the new compounds in patents. This could be done quite easily and at moderate cost.

Thirdly, collaboration between Chemical Abstracts Service and Derwent should be encouraged, since one covers individual compounds in patents and the other indexes them generically.

Fourthly, CAS should itself develop a system for substructure searching. This has, of course, been under study using various approaches for a number of years, but a working system at a realistic cost is probably still several years away.

Let us examine these possibilities in more detail.

IDC has been in existence for about a decade, yet it has attracted few members outside its three founder companies, Bayer, BASF, and Hoechst. Undoubtedly, this is due very largely to the high cost, which IDC fully admits to. This in turn is because of the care with which documents are analyzed, and bibliographic details, especially of patents and patent equivalents, are checked. However, the encoding itself has been largely mechanized, and, for example, GREMAS fragment codes are generated automatically from CA connection tables. If, however, we are looking to on-line searching, the hierarchical nature of the GREMAS code and the heavy postings of some terms make this difficult to implement.

The second proposal, to cipher all compounds exemplified in patents in WLN, would be welcomed by the substantial number of companies who already use WLN for indexing their own collections of compounds. The AIOPI study group estimated that in Farmdoc plus Agdoc there were rather more than 200 000 new compounds a year.

At the present time on-line search systems for WLN are limited to in-house collections; the one in operation for the ICI Compound Collection provides bit and string searching and the ability to display the number of hits and the WLN ciphers and registry numbers of compounds. Atom-by-atom searching and output to display structures is provided by overnight batch operation. A British software house is, however, studying the requirements for a fully interactive system.

The third possibility, for closer collaboration between CAS and Derwent, was discussed at the first meeting of the newly formed Derwent Chemical Coding Committee. It was agreed that addition of Registry Numbers of compounds indexed by CAS to the Derwent data base would be useful and should be explored. It would provide strong links between two major systems, but it would not directly assist substructure searching.

An on-line alternative, which is now becoming available through SDC in this country and Infoline in the United Kingdom, is for substructure search on CA systematic names. This will clearly be useful for compounds containing a definitive ring system and for the simpler substituents. It will also be useful where specific patterns of substitution are being sought, but for functions like esters or amidines which can serve as links between two larger entities, the same problem arises as with notation. The exact way in which a group is described depends on the nature and relative seniority of the parent compound and its attachments.

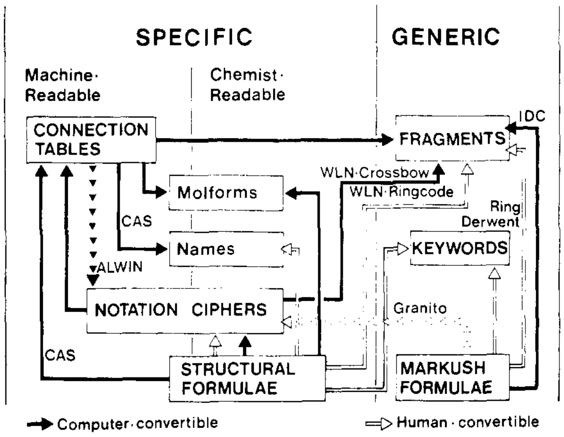


Figure 2. Relationships and interconvertibility of structural descriptors.

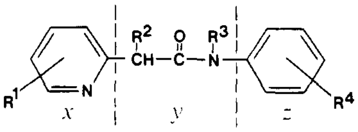
Viewing the situation in another way, Figure 2 shows the categories and relationships for the various types of descriptors of molecular architecture arranged in three columns. The first is Specific and Machine-readable, and it contains two members: connection tables and notation ciphers. Both of these give a complete description of structure in a form suitable for computer processing. Ciphers extend also into the second column, Specific and Chemist-readable, and this also includes molecular formulas, systematic names, and structure formulas. The third column is headed Generic and it contains keywords, fragments, and Markush formulas. In moving from left to right we are progressing from a strictly logical, analytical, and microscopic mode of description to a looser and more empirical type of macroscopic description. This enables us to designate larger entities which are recognized to be of practical importance.

A number of interconversions are possible between these types of record; some can already be carried out by computer, while others are done manually at the present time. A broken line denotes a conversion that is either possible or has not yet been completely described. As far as specific descriptors are concerned, there is no doubt that we wish to move increasingly toward fully automated processing, because this not only saves human effort but also eliminates the variability in its performance.

If we were starting afresh today in the situation of 20 years ago, we would recognize the wisdom of developing fragment codes and notation systems in a strictly algorithmic way, so that they could be generated entirely by computer; that is, they would be logical rather than empirical. The area where this seems to apply most strongly is in development of improved coding for Derwent CPI. It is particularly noteworthy that Derwent is still entirely dependent on intellectual input, whereas IDC has mechanized the GREMAS coding of Markush structures. The generation of Ringcodes from WLN ciphers has also been mechanized to a considerable extent.

Fragment codes still have a place in providing a preliminary bit screen for rapid searching of large files, but if they are to be useful, they need to have chemical significance. In contrast to Derwent, it appears that part of the reason why CAS does not yet have a practical substructure search system is not only because its files are so large, but because the fragments generated from connection tables have not related sufficiently closely to the types of substructures and relationships among them which chemists require. The set of fragments for computer generation needs to be chosen judiciously with the emphasis on chemical significance.

In my opinion a notation system still has an important role to play in this area, even though it may find a diminishing role outside the computer. Most systems of notation aim at expressing customary units, like rings and functions, in recognizable forms. They therefore complement connection tables, which are better able to deal with the requirement of a unique descriptor for each structure. The two are related and interconversion is already possible. Although we are moving toward direct structure input and output on the screen of a terminal, behind the scenes an algorithmic notation could still provide the metalanguage for economical processing of substructure searches. In the CROSSBOW system as it operates in ICI, we can key in a cipher on-line, and, having found that it is new to the Company Compound File, we can register it with immediate generation of fragments, so it becomes available for bit fragment search and string search



R^1 = Cl, OMe or CF_3
a b c

R^2 = H, Me or Cl
h d a

R^3 = H, Me or Et
h d e

R^4 = H, Cl, OMe or NO_2
h a b f

$(x + (a,b,c)) + (y + (a,d,h) + (d,e,h)) + (z + (a,b,f,h))$

Figure 3. A generic formula as a Boolean expression.

of the notation. These two methods suffice for the great majority of our searches.

At present we think of a notation mainly as a means of encoding individual compounds, but if we are aiming at a comprehensive system, we need one which can deal equally well with Markush formulas or similar generic descriptions. Some work on these lines has already been done by Granito with WLN, although details have not been published. It is an approach which has come up for discussion several times in Britain, and we were pleased to learn recently that Krishnamurthy, author of the ALWIN notation,[2] has developed a form able to deal with generic structures.

A further point, which seems not to have been widely recognized, is the similarity between a Markush structure and a nested Boolean expression. Figure 3 provides a simple illustration. The essential structure comprises a pyridine ring and a benzene ring linked by a chain including an amide group. These three components may be designated *x*, *y* and *z*, and the optional substituents may be designated by the letters *a* to *h*. The whole formula may then be reduced to a nested Boolean type of expression. This shows not only which sets are present, but also specifies their relationships. Although the various sets denoted by R^1 to R^4 have members in common, the expression shows clearly which ones may be present at each point in the main structure.

In this respect it resembles the cipher of a notation, and from the viewpoint of searching a patent data base there would clearly be immense advantages in having a system which preserved this structure of relationships in addition to providing a precise description of the individual components. It reinforces the conclusion that notation has a place in our future chemical information systems and that this notation should take into account the description of generic structures as well as specific ones.

Such an approach could prove useful for the journal literature as well as patents. Many papers describe series of related compounds; if these were reduced to a generic formula of a type suitable for computer processing, we could handle both the journal and the patent literature in a new way. For each paper there would be generated one or a few generic formulas which embraced all the compounds described in it. This would drastically reduce the size of files and would thereby make a major contribution to solving the problem of searching large files. Although it should not be expected that a generic formula would be adequate as a basis for all types of search request, our present experience of actual enquiries indicates that bit fragment and string search would suffice for the majority of them and, moreover, provide the inexpensive preliminary screens which must precede search of connection tables of large files.

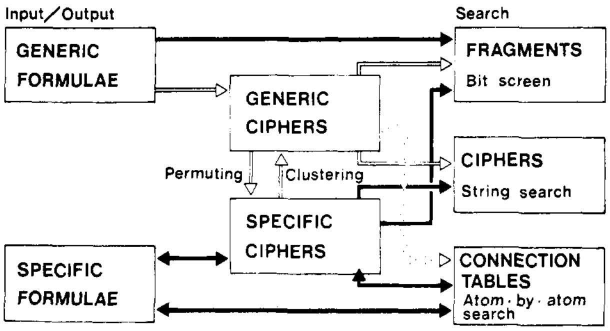


Figure 4. Features of an algorithmic notation.

By way of illustration, it is worth noting an approach which in principle is already possible. This might be described as cluster analysis of CA names for compounds occurring in the same patent or paper. A perusal of the CA Registry Number Indexes shows clearly that blocks of consecutive registry numbers relate to compounds having similar names, and therefore presumably occurring in the same paper. Although this amounts to a purely mechanical clustering of related compounds, it appears to support the conclusion that there could be great benefit to chemical information retrieval arising from a notation able to provide generic descriptions of molecular structures and hence a concise means of designating a group of compounds.

CONCLUSION

With an algorithmic notation (see Figure 4), ciphers could be generated either from connection tables or from structural diagrams composed on the screen of a terminal. Equally, they could be composed by trained encoders in the same way WLN ciphers are now; this is still a very economical means of entering structures into a computer.

REFERENCES AND NOTES

(1) A. W. Nineham, "Comparison of Ringdoc and Farmdoc Codes", Derwent International Patents Conference Proceedings, 1978, p 340.

(2) E. V. Krishnamurthy, P. V. Sankar, and S. Krishman, "ALWIN-Algorithmic Wiswesser Notation System for Organic Compounds", *J. Chem. Doc.*, **14**, 130 (1974).