A Comparison of Different Approaches to Markush Structure Handling[†]

JOHN M. BARNARD

Barnard Chemical Information Ltd., 46 Uppergate Road, Stannington, Sheffield S6 6BX, U.K.

Received November 17, 1990

The history of the development of computer systems for storage and retrieval of Markush structures is reviewed briefly. The systems currently being developed by Chemical Abstracts Service, Derwent Publications/Questel/INPI, and International Documentation Company for Chemistry (IDC) are introduced, and the similarities and differences between the approaches they use for input representation and search are examined, especially in relation to the handling of generic nomenclature. the prospects for future development of such systems is discussed in light of recent and continuing research work, as is the potential for exchange of Markush structure databases between the different search systems.

HISTORY

The history of systems for Markush structure handling extends back to the earliest days of computerized chemical information systems.[1] Most of the systems developed in the early 1960s employed fragment code representations of the chemical structures and involved manual encoding of the patent documents. Structural features were represented by particular fragments, and the structures were originally searched using punched-card equipment. Among the best-known of such systems are Derwent's CPI code[2] and the IFI/Plenum system.[3]

Another system, less widely known, is the GREMAS code, a very sophisticated fragment code, developed in the late 1950s by Robert Fugmann at Hoechst.[4] It is complicated to learn to use and is expensive to encode manually and difficult to search. Recently, some work has been done at Hoechst on a program called GREDIA,[5] which allows graphical input of queries, with automatic generation of the appropriate GREMAS search strategy, in the manner of Derwent's TOPFRAG program;[6] there are plans to develop a microcomputer equivalent to GREDIA. A number of studies[7,8] have shown the retrieval performance of GREMAS to be markedly superior to that of other systems available at present.

Line notations were looked at briefly in the late 1960s as a possible representation for Markush structures. A joint project between the National Bureau of Standards and the U.S. Patent and Trademark Office, which had also looked at a variety of connection-table formats for generic structures,[9] proposed some extensions to the Hayward notation;[10] one or two groups also tried extending the more popular Wiswesser notation, and Dyson suggested extensions to his own IUPAC notation.[11] In 1979 Lynch and Krishnamurthy[12] examined Krishnamurthy's ALWIN notation,[13] which looked more promising, but eventually also proved unsatisfactory, though some ideas from their work influenced Lynch's subsequent research on Markush structure retrieval.

One early system which deserves particular mention is that developed by Meyer and his colleagues at BASF.[14] This is a fully topological, connection-table-based system for storage and retrieval of Markush formulas; work on it was started as early as 1958, and it was fully operational by 1966. It is, of course, very restricted in the types of structure it can handle, but it was nonetheless years, if not decades, ahead of its time.

In 1979 a major, and continuing, research project at Sheffield University began under the direction of Professor Lynch.[15-25] Commercial interest in this project is indicated by the fact that during its course it has received financial support from Chemical Abstracts Service, Derwent Publications, IDC International Documentation Company for Chemistry, and Questel SA, as well as from U.K. government sources.

Since the mid-1980s three separate groups have been developing operational systems for topological storage and retrieval of Markush structures.[26] Markush DARC is a joint development by Derwent Publications Ltd (U.K.), Questel SA (France), and INPI (the French Patent Office). It is based on extensions to the DARC system for substructure searching of specific structure databases.[27] MARPAT is being developed by Chemical Abstracts Service and utilizes some of the principles behind the STN substructure search system, as well as much new work.[28] The GENSAL/GREMAS system is being developed by the International Documentation Company for Chemistry (IDC). IDC is an independent company, based near Frankfurt, but its shareholders are the major German and Austrian chemical and pharmaceutical companies. It provides very high quality information systems (such as the GREMAS fragment code) to its member companies only. The GENSAL/GREMAS system is strongly based on the research work done at Sheffield and incorporates software developed at the University; it is also based on the existing GREMAS fragment code.[29]

INPUT REPRESENTATION

Both Markush DARC and MARPAT have taken a similar approach to the input of Markush structures from patents, enhancing the existing generic query capabilities of their respective substructure search systems for specific structures.

This means that a structure diagram is input for the invariant or core part of the structure, which includes variable groups (R groups, G groups, etc.). Each of these is then defined by means of a series of alternative structure diagrams, which can themselves contain further variable groups. Though the systems are almost entirely graphics-based, some shortcuts are available for common groups; MARPAT also has special "SO" (substitution optional) and "SR" (substitution required) indicators, which can be applied to individual values for the variable groups. Both systems have various limits on the number of variables, the number of attachment points each can have, the number of alternative values for each, or the total number of atoms over all the alternatives.

One interesting point of difference between the two is the handling of variable-position attachment. In Markush DARC the variable-position group is placed next to the ring, and the user indicates which positions are possible for the attachment. This is expanded internally into separate structure diagrams for each alternative, each with a fixed attachment position

[†] This paper was presented (under the title Technical Aspects of Markush Structure Handling) at the 200th National Meeting of the American Chemical Society, Washington, DC, August 1990.

© 1991 American Chemical Society

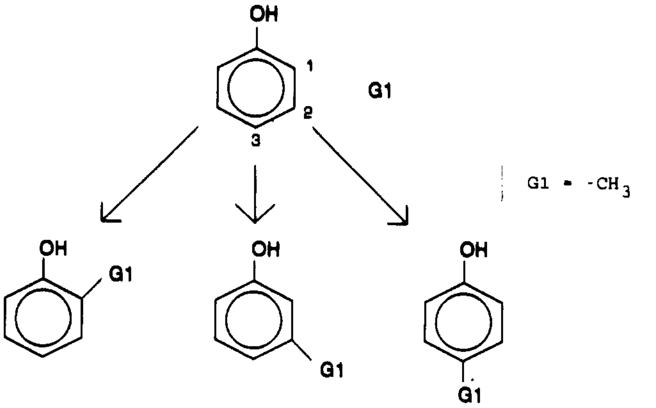


Figure 1. Input of a variable-position group in Markush DARC. The single input structure is expanded internally into the three possible fixed-position attachments.

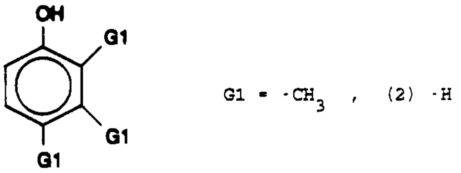
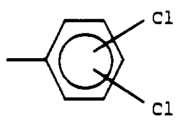


Figure 2. Input of a variable-position group in MARPAT. It is stipulated that two of the occurrences of G1 are hydrogens.

R1 = SD



/ pyridyl [2] SB <1-2> Cl

Figure 3. Example of the definition of a R group in GENSAL (explanation in text).

Table I. Superatoms Used in Markush DARC

CHK	alkyl, alkylene
CHE	alkenyl, alkenylene
CHY	alkynyl, alkynylene
ARY	aromatic carbocyclic system
CYC	cycloaliphatic system
HEA	heteroaromatic monocycle
HET	nonaromatic hetero monocycle
HEF	fused heterocycle
MX	any metal
AMX	alkali/alkaline earth metal
A35	group IIIA-VA metal
TRM	transition metal
LAN	lanthanides
ACT	actinides
HAL	halogen
ACY	acyl
DYE	dye group residue
PEG	polymer end group
POL	polymer or polypeptide residue
PRT	protecting group
XX	any atom or group (except hydrogen)
UNK	undefined group

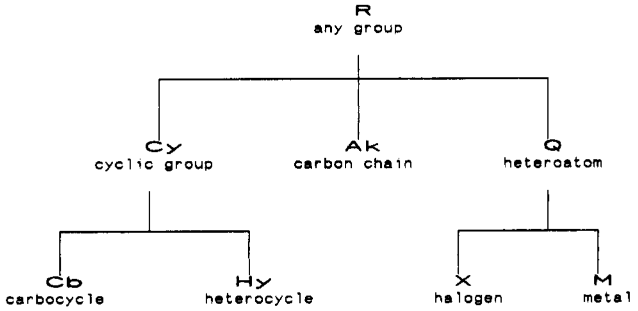


Figure 4. Hierarchical Generic Groups (HGGs) used in input of structures to MARPAT.

(Figure 1). In MARPAT, the input structure is shown with the G group attached separately at each possible attachment point, and the definitions then specify that all but one of the G groups is hydrogen (Figure 2).

The GENSAL/GREMAS system uses a rather different approach, in which GENSAL is used as the input representation. GENSAL[16] (which is an acronym for GENeric Structure LAnguage) is a formal language, like a programming language, and it includes both structure diagrams and chemical nomenclature, as well as special operators, such as "SB" (substituted by). Figure 3 shows the GENSAL definition of a variable group, with two alternatives. The first is a structure diagram, and the second uses nomenclatural terms to show that R1 can be a pyridyl group (attached in its 2-position) which is substituted by 1 or 2 chlorines. The diagram also illustrates that variable-position attachment can be shown by means of the convention of a bond drawn to the middle of a ring.

GENSAL is designed to be as close as possible to the language of patent specifications, and research at Sheffield is currently looking at the possibility of generating GENSAL at least semiautomatically from machine-readable patent specifications.

IDC has a microcomputer-based input program, called MicroGensip, which incorporates specially adapted versions of the GENSAL Interpreter program originally written at Sheffield,[19] and the PSIGEN chemical structure graphics systems developed by Hampden Data Services.[30] As far as possible, there are no fixed limits on the size of the GENSAL or of the internal representation derived from it.[18] The limit is essentially one of available memory, though this has proved to be a problem on MS-DOS microcomputers, and IDC is currently considering reimplementation under a different operating system.

REPRESENTATION OF GENERIC GROUPS

One of the major problems with Markush structures, beside their inherent variability and complexity, is the representation of generic groups and the matching of these against specific examples of them. A generic group can be considered as one which cannot be shown by a single atom-bond connection table, and examples include expressions like "alkyl", "heterocyclic", and "cycloalkyl". In some cases the expression may have no direct association with structural features, like "electron-withdrawing group".

Each of the three systems uses a different approach to representing these groups.

Markush DARC uses a set of 22 *superatoms*, each representing a different type of group. The available superatoms are listed in Table I. A superatom can be treated as a normal atom, but is able to represent a whole group, though some represent only single atoms (e.g., a transition metal atom). Some superatoms can have special "attributes" applied to them (e.g., "LO", "MID", and "HI" for the number of carbon atoms in an alkyl chain represented by the CHK superatom), and additional textual descriptors can be used to give more detailed information. There is no relationship defined between the different superatoms, all of which rank equally in a "flat" hierarchy.

MARPAT uses an approach which at first appears similar to Markush DARC's, but in fact probably has at least as much in common with the GENSAL approach. As with superatoms, the *hierarchical generic group* (HGG) nodes can be treated for many purposes like "real" atoms, but can represent whole groups. There are only 8 HGGs that can currently be used

Table II. Attributes Which May Be Applied to Generic Groups in the MARPAT System

GG	number and type of HGG groups in group
CA	number and type of acyclic atoms
RC	number of rings
RS	size of rings
RA	number and type of ring atoms
EC	number and type of elements
BD	number and type of bonds
DC	number and type of connectivities for atoms
FA	number and type ring fusion atoms
CH	collective charge on a group
AN	number and kind of atoms attached outside the group
SS	real-atom substructure contained in the group
TX	text qualifier for the group

Table III. Classes of Parameters Used in the GENSAL/GREMAS System

GREMAS district (number of rings, chains, etc.)
atom counts (number of atoms of different elements)
unsaturation (number of double/triple bonds)
carbon chain branching (number and type of branch points)
ring fusion (number and type of ring fusion atoms)
charges/radicals (number and value)
connections to "parent" structure (number and bond orders)
connections to "child" structure (i.e., substitution pattern)

in file structures (though in many cases the actual nomenclatural terms, such as "alkyl" can be used for input and display), and unlike superatoms they are arranged in a strict hierarchy of increasing specificity, as shown in Figure 4. A set of numerical attributes can be applied to each HGG, describing certain structural features. These attributes, which are listed in Table II, are very similar in concept to GENSAL's parameter lists discussed below. There are also qualitative categories of HGG, called generic group categories, which are rather similar to Markush DARC's superatom attributes.

GENSAL uses ordinary nomenclatural terms such as "alkyl" to show these groups, and they can be qualified by means of a set of structural parameters, each given a numerical value or range of values. The internal representation is based entirely on the structural parameters. The principle behind the use of parameter lists is that the values given for each parameter are sufficient to define the structural characteristics of the group in question. For example, an alkyl group has one chain of atoms, which may have any number of branches; has any number of carbon atoms, but no heteroatoms; has no double or triple bonds; and has no rings. This is effectively a complete definition of an alkyl group. GENSAL parameter lists can be likened to applying the MARPAT attributes to the most generic HGG, "R". The parameters contain the information that would otherwise be given by the actual HGG node used.

In the original work at Sheffield[17] a set of 12 parameters was used, but for reasons connected with the use they intend to make of them, IDC has developed a more comprehensive set of 37 parameters,[29] in 8 classes, which are listeed in Table III, though they remain generally applicable. Internally, the generic structure is represented as an AND/OR logic tree (the ECTR[18]) in which each node, called a *partial structure*, may be a partial connection table or a parameter list. Textual expressions can also be applied to generic partial structures.

SEARCHING

In both Markush DARC and MARPAT, the searching is based on the software originally developed for generic queries in specific structure databases. That is, there is a screening search based on limited-environment fragments, followed by an iterative atom-by-atom search on those structures which pass the screening stage.

The nature of the fragments used for the screening stage differs between the two systems (FRELs in Markush DARC, augmented atoms, atom sequences, bond sequences, etc. in MARPAT), and of course the Markush systems include the superatoms and HGGs, which do not occur in the corresponding specific structure systems. The fragment search obviously requires the logical relationships inherent in the generic structure to be taken into account.

At the atom-by-atom stage, the alternative groups possible mean that much more backtracking is required than in specific structures, because there are more potential matches to test. This means that atom-by-atom search times are correspondingly slower.

The GENSAL/GREMAS system does things rather differently. It is at present, at least, a hybrid or half-way-house system. Programs have been developed which can generate GREMAS fragments automatically from the internal representation derived from GENSAL input, and operational use of these is expected to start in the near future, which will allow manual GREMAS coding to be phased out, at least for the majority of structures. Automatic GREMAS code generation has been done for many years for specific structures, and in fact the whole Chemical Abstracts Registry File is searched by GREMAS code at IDC. When GREMAS fragments have been generated for the Markush structures, the file can then be searched in exactly the same way as it has been since the system started in 1959.

In the longer term it is hoped to develop a more sophisticated search system, including an atom-by-atom search stage, and this will be based on continuing research work at Sheffield. However, the GREMAS search will probably be retained as a screen, and to allow continuing search of the backfile which now contains over 9 million structures.

GENERIC-SPECIFIC MATCHING

In Markush structure searching one of the most vital things is the ability to match a specific group (described by a structure diagram or connection table) against a corresponding generic group. For example, to match a group described as "alkyl" against an *n*-butyl group.

In Markush DARC this is not possible at all at present; superatoms cannot be matched against real atoms. In formulating queries it is necessary to search, in addition to the relevant superatom, for all specific groups to which it could correspond in order to achieve complete recall. This is obviously a very serious limitation, and it substantially restricts the usefulness of the system. Indeed, in their present form, the superatoms are able to function as little more than a very crude fragment code, involving only 21 fragments. However, the problems caused by these limitations are well recognized, and work is in progress to allow "translation" between superatoms and real atoms in future versions of the system.

MARPAT, on the other hand, has such translation already. Two basic requirements can be identified for the matching of a generically described group against a specifically described one:

(a) The representations must be converted to descriptions which can be compared directly.

(b) The groups to be compared must have common boundaries.

In the case of the first requirement it is fairly obvious that comparison must be done at the level of the generic description, since the automatic generation of all specific embodiments of the generic group will be unfeasible in all but the most trivial of cases. Thus, in order to carry out the comparison in MARPAT, the various real-atom groups are automatically converted to the corresponding HGGs. In fact an expanded hierarchy of HGGs is used in the comparison, involving 11 different groups.[28] This is the point where the design of the generic group hierarchy becomes so important, because in choosing the set of real atoms to map onto a single HGG (i.e., in identifying the boundaries between the groups to be compared), the definitions of HGGs ensure that the boundaries between the segments of the structure can be determined reproducibly, thus providing a solution to the second requirement. In practice, of course, this becomes quite complicated and involves "shifting" the boundaries between the constant and variable parts of the structure in order to ensure that they correspond to the boundaries between HGGs. At present, MARPAT matches only on HGG type and does not compare the attributes applied to each. This facility will be added later, though the generic group categories are already searchable.

In the GENSAL/GREMAS system, the GREMAS terms can be derived from the parameter lists just as they can be from connection tables, and the matching is then done at the level of GREMAS fragments. The enhanced set of parameters used by IDC in their system was chosen partly with GREMAS generation in mind. The GREMAS district parameter is in fact a way of chopping up the structure reproducibly; indeed, though they are not the same there are definite parallels between the definition of GREMAS districts and the definition of HGGs.

There are a number of ideas that are being worked on in Sheffield for using the parameters directly in matching generic and specific groups. Again, these involve chopping up the generic structure reproducibly into separate units, in this case forming what is called a reduced graph,[22] in which each node represents a set of atoms. There are several ways in which the structure can be chopped up, though most of the work at Sheffield has been based on putting the boundaries between the chains and the rings (thus paralleling the division between the Cy and Ak/Q HGGs). For each separate unit (reduced graph node) which contains real atoms, a parameter list can be derived simply by counting the occurrences of features described by each parameter. Matching can then be achieved by comparing parameter lists for overlap. Some related ideas have also been studied in Japan. The extended block-cutpoint tree developed by Nakayama and Fujiwara[31] is also a form of reduced graph, and Tokizane et al.[32] have proposed a matching algorithm for generic group attributes which are analogous to MARPAT's HGGs and GENSAL's parameter lists.

CONCLUSIONS

Three different systems are being developed, each with its own peculiarities of generic structure representation and search. Each organization is also building its own database; indeed two different databases are being built for search with the Markush DARC system, one by Derwent and one by INPI.

None of the systems under development appears to have a total monopoly of good ideas, and the competitive development of a number of different search systems is likely to result in better systems in the end than if there was just a single organization working on it. However, the cost of development of the systems, high though it is, is likely to be dwarfed by the ongoing cost of building the database, and from the users' point of view there would seem to be little advantage in three different organizations separately encoding the same information from the same patents in slightly different ways.

In principle, there does seem to be some hope that the input representations could be interconverted in some way, probably at the connection-table level. The greatest problem would lie with generic groups, though Tables II and III show that there is certainly sufficient common ground between MARPAT's HGG attributes and GENSAL's parameter lists for some sort of interconversion to be fairly readily achievable. The attributes used with the Markush DARC superatoms have a lower level of specificity, though again there does seem to be a basis for a mapping (albeit perhaps with some information loss) between them and the other representations; the additional textual descriptors which can be applied to superatoms may also have a part to play in interconversion, as they are intended to in translation between superatom and real-atom groups.

The greatest problem in any sort of shared database is likely to be agreeing on the depth to which patents should be indexed. GENSAL certainly seems to allow much more detailed description of generic structures, though not all the information it can represent can be used at the search time, at least at present. This reflects IDC's requirement for the best possible system that technology will allow, and IDC will certainly not be happy with a topological system whose retrieval performance falls below that already achieved by the GREMAS fragment code.[8] However, the more deeply a patent is indexed, the more expensive database creation becomes, and this is the point at which technical considerations have to take second place to commercial and economic ones.

ACKNOWLEDGMENT

I am grateful to the following, who have kindly provided information on the systems they are developing and commented on drafts of this paper: Tommy Ebe and Patricia Wilson (Chemical Abstracts Service), Tony Ferns and Kathie Shenton (Derwent Publications Ltd.), Arthur Kolb and Günter Stiegler (IDC), Michael O'Hara and Osman Sultan (Questel), and Geoff Downs, Valerie Gillet, Michael Lynch, and John Holliday (Sheffield University).

REFERENCES AND NOTES

(1) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Generic structure storage and retrieval. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 264–270.

(2) Simmons, E. S. Central Patents Index chemical code: a user's viewpoint. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 10–15.

(3) Kaback, S. M. The IFI/Plenum chemical indexing system. In *Computer handling of generic chemical structures*, Proceedings of a conference organized by the Chemical Structure Association, University of Sheffield, England, March 26–29, 1984; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 49–65.

(4) Rössler, S.; Kolb, A. G. The GREMAS system, an integral part of the IDC system for chemical documentation. *J. Chem. Doc.* **1970**, *10*, 128–134.

(5) Fricke, C.; Nickelsen, I.; Fugmann, R.; Sander, J. GREDIA: a new access to GREMAS databases. *Tetrahedron Comput. Methodol.* **1989**, *2*, 167–175.

(6) Meyer, D. E. Special-application software for chemical structures. In *Chemical structure software for personal computers*; Meyer, D. E., Warr, W. A., Love, R. A., Eds.; American Chemical Society: Washington, DC, 1988, pp 73–81.

(7) Suhr, C.; von Harsdorf, E.; Dethlefsen, W. Derwent's CPI and IDC's GREMAS: remarks on their relative retrieval power with regard to Markush structures. In *Computer handling of generic chemical structures*, Proceedings of a conference organized by the Chemical Structure Association, University of Sheffield, England, March 26–29, 1984; Barnard, J. M., Ed.; Aldershot: Gower, 1984; pp 96–105.

(8) Schoch-Grübler, U. (Sub)structure searches in databases containing generic chemical structure representations. *Online Rev.* **1990**, *14* (2), 95–108.

(9) Hayward, H. W.; Tauber, S. J. The HAYSTAQ experiment. In *Proceedings of the Fifth Annual Meeting of the Committee for International Cooperation among Examining Patent Offices (ICIREPAT)*, London, Sept 1965; Thompson: Washington, 1966; pp 337–350.

(10) Sneed, H. M. S.; Turnipseed, J. H.; Turpin, R. A. A line-formula notation system for Markush structures. *J. Chem. Doc.* **1968**, *8*, 173–178.

(11) Dyson, G. M. Generic (or Markush) groups in notation and search programs, with particular reference to patents. *Inf. Storage Retr.* **1964**, *2*, 59–71.

(12) Krishnamurthy, E. V.; Lynch, M. F. Analysis and coding of generic chemical formulae in chemical patents. *J. Inf. Sci.* **1981**, *3*, 75–79.

(13) Krishnamurthy, E. V.; Sankar, P. V.; Krishnan, S. ALWIN—Algorithmic Wiswesser notation system for organic compounds. *J. Chem. Doc.* **1974**, *14*, 130–141.

(14) Meyer, E.; Schilling, P.; Sens, E. Experiences with input, translation and search in files containing Markush formulae. In *Computer handling of generic chemical structures*, Proceedings of a conference organized by the Chemical Structure Association, University of Sheffield, England, March 26-29, 1984; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 82-95.

(15) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Computer storage and retrieval of generic chemical structures in patents. 1. Introduction and general strategy. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148-151.

(16) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer storage and retrieval of generic chemical structures in patents. 2. GENSAL, a formal language for the description of generic chemical structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151-161.

(17) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and retrieval of generic chemical structures in patents. 3. Chemical grammars and their role in the manipulation of chemical structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 161-168.

(18) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer storage and retrieval of generic chemical structures in patents. 4. An extended connection table representation for generic structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160-164.

(19) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer storage and retrieval of generic chemical structures in patents. 5. Algorithmic generation of fragment descriptors for generic structure screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57-66.

(20) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer storage and retrieval of generic chemical structures in patents. 6. An interpreter program for the generic structure description language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66-71.

(21) Gillet, V. J.; Welford, S. M.; Lynch, M. F.; Willett, P.; Barnard, J. M.; Downs, G. M.; Manson, G.; Thompson, J. Computer storage and retrieval of generic chemical structures in patents. 7. Parallel simulation of a relaxation algorithm for chemical substructure search. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118-126.

(22) Gillet, V. J.; Downs, G. M.; Ling, A. (B.); Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs, and their application in generic chemical structure retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126-137.

(23) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer storage and retrieval of generic chemical structures in patents. 9. An algorithm to find the Extended Set of Smallest Rings (ESSR) in structurally explicit generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207-214.

(24) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer storage and retrieval of generic chemical structures in patents. 10. The assignment and logical bubble-up of ring screens for structurally explicit generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215-224.

(25) Lynch, M. F. Generic chemical structures in patents (Markush structures): the research project at the University of Sheffield. *World Patent Inf.* **1986**, *8*, 85-91.

(26) Barnard, J. M. Online graphical searching of Markush structures in patents. *Database* **1987**, *10* (3), 27-34.

(27) Shenton, K. E.; Norton, P.; Ferns, E. A. Generic searching of patent information. In *Chemical structures: the international language of chemistry*, Proceedings of an international conference at the Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, May 31-June 4, 1987; Warr, W. A., Ed.; Springer: Heidelberg, 1988; pp 169-178.

(28) Fisanick, W. The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145-154.

(29) Stiegler, G.; Maier, B.; Lenz, H. Automatic translation of GENSAL representations of Markush structures into GREMAS fragment codes at IDC. In *Proceedings of the 2nd International Conference on Chemical Information Systems*, Noordwijkerhout, The Netherlands, June 1990; Warr, W. A., Ed.; Springer: Heidelberg, in press.

(30) Love, R. A. Structure drawing software. In *Chemical structure software for personal computers*; Meyer, D. E., Warr, W. A., Love, R. A., Eds.; American Chemical Society: Washington, DC, 1988; pp 9-36.

(31) Nakayama, T.; Fujiwara, Y. Computer representation of generic chemical structures by an extended block-cutpoint tree. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 80-87.

(32) Tokizane, S.; Monjoh, T.; Chihara, H. Computer storage and retrieval of generic chemical structures using structure attributes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 177-187.

——1990 HERMAN SKOLNIK AWARD PAPER——

Computer Representation and Handling of Structures: Retrospect and Prospects[†]

ERNST MEYER[‡]

Friedelsheimer Strasse 18, D-6700 Ludwigshafen, FRG

Received October 3, 1990

Topological encoding of structures was a necessary supplement to documentation methods. As a practical approach, it was developed first for substructure retrieval in chemical formulas, but it also proved useful in other areas such as reaction retrieval, synthesis planning, semantical and syntactical concept interrelations, patent claims examination, drug design, and even in other disciplines like electrical and mechanical engineering. A survey of three decades of methodological development is given, and some newer trends are indicated.

INTRODUCTION

It becomes more and more difficult to retain an overview of our growing treasure of knowledge, even in a partial area. Documentation methods were developed long before the advent of computers, but indexes, classifications, card files, and similar tools were soon not effective enough. The computer, invented just in time, was at first able to help with the multidimensional search for words and classes, but as a result of the knowledge explosion even index and full-text searches soon became insufficient means, and, especially in chemistry, classification systems became steadily more effort consuming and could not keep pace with the rapid appearance of new concepts and requirements.

Mankind can increase his ability immensely by protheses, tools, and machines. But if one wants to use the computer as a thinking machine, it has to be able to handle not only numbers and character strings but also structures, because our thinking—especially in organic chemistry—proceeds in structures. This need appeared in chemistry very early due to the size and long-life of its treasure of knowledge. Fortunately, chemical structural formulas were quite suitable models for the development of useful computer methods. The approach was supplied by an old branch of mathematics: topology or—more exactly—graph theory.

TOPOLOGY

Graph theory reduces a structure to a set of nodes and the connecting edges. Numbers can be given to both types of elements, and attributes (consisting of words and/or numbers) can be attached to each node or edge in order to characterize them. In this way it becomes possible to localize a substructure in filed structures. It was an American mathematician from Cambridge, MA, Calvin N. Mooers, who suggested in 1951[1] recording chemical structural formulas in this manner on a computer for structure and substructure searches. However, he never practiced this approach himself.

[†] This paper was presented at the ACS Symposium on Topological Storage of Structures in Chemistry and Related Fields, Washington, DC, Aug 28, 1990.

[‡] Formerly with BASF Aktiengesellschaft.

0095-2338/91/1631-0068$02.50/0 © 1991 American Chemical Society