The Sheffield Generic Structures Project—a Retrospective Review

Michael F. Lynch* and John D. Holliday

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received November 13, 1995[⊗]

The problems posed by the requirements for storage and manipulation of generic chemical structure definitions in patents are reviewed. Chemists and patents agents have developed an armory of linguistic devices over many decades so that a generic structure description can describe large and often unlimited numbers of substances as a result of the combinatorial opportunities provided. The nature of these linguistic devices is defined, and the theoretical foundations devised during the Sheffield project for the successful solution of the problems in order to provide the desired retrieval facilities are reviewed. Progress toward the practical implementation of a system based on these solutions is evaluated. The relevance of the data structures and algorithms devised in this work to the issues raised by developments in combinatorial libraries is also reviewed.

1. INTRODUCTION

In 1978 a report by a specialist group constituted by the British Library R&D Department[1] reported on issues and requirements in chemical information services and provided a welcome pointer to what might be termed, at that time, the major remaining issue in classical chemical structure information and retrieval systems, that of generic chemical structures. The report pointed out the need for more adequate access to the information in chemical patents, which takes the form of generic chemical structures. These are very important for the fine chemicals industry and are also known as Markush structures after the inventor who won a law suit against the U.S. Patent Office on the issue of how such structures could be described in patents.

A number of information services provided access to patent information at that time; the dominant service was the World Patent Index (WPI) of Derwent Publications Ltd., but International Documentation in Chemistry mbH, then based in Frankfurt, Germany, IFI Plenum, and the American Petroleum Institute (API) also provided services to industry which included access to the structural information in patents in greater or lesser measure.[2] Some chemical companies, most notably those in Germany, also operated internal services, of which that of BASF, Ludwigshafen, already in operation as early as the 1960s, and designed by Dr. Ernst Meyer, was the most advanced in design.[2−4]

With the exception of the BASF in-house system, the services operated exclusively on the basis of fragmentation codes, i.e., the generic structure information was summarized as a series of fragments, derived from the structure descriptions by skilled analysts. Without going into detail, there were several contrasting bases for the design of these codes. That used by Derwent Publications Ltd., the Derwent Chemical Code, was based on experience with patents and therefore on pragmatically determined need.[5] Since the concerns reflected in industrial chemistry change with time, the code also needed to change. At the very least, this constituted no small inconvenience for the searcher when devising queries, since these needed to reflect different time-slices. Again, the characteristics of the code were such that the performance levels attained with it, in terms of recall and precision, left much to be desired.[5] The IDC code, in contrast, was based on a systematic and rational design, that of Dr. Robert Fugman.[6] This was the GREMAS code, which was little understood for many years outside of Germany. It had a strong structural and hierarchical basis to its design and was logical to such a degree that it was later possible to generate it automatically not only from representations of specific substances but also from generic structures.[7]

The HAYSTAQ project at the U.S. Patent Office and the National Bureau of Standards had earlier attempted to address the complexities of generic chemical structures by devising a topological representation as early as 1958, as described by Koller, Marden, and Pfeffer.[8] At that time, the computing facilities available to them were paltry and lacking in the storage space and processing power needed to achieve viable solutions.

In early 1979, a British Library R&D Department Fellowship enabled Professor E. V. Krishnamurthy of the Indian Institute of Science at Bangalore, India, to visit Sheffield to tackle the problems with Lynch. They suggested that an approach based on formal grammar theory might be appropriate for the representation of the structures as well as holding some prospect of providing search algorithms.[9] Work on these problems began with Barnard and Welford later in 1979. This progressed well; the British Library R&D Department funded continuation of the initial doctoral researchers from 1981 until 1983, at which point finances became available from both Chemical Abstracts Service and from Derwent Publications Ltd., with support from the latter continuing for several years. A series of publications documented progress in the project.[10−32]

In 1984, International Documentation for Chemistry mbH, the consortium of German chemical companies, began support for the research which continued until 1992, when a licence for the implementation of a database construction system using the Sheffield methods and software was negotiated. At that stage, the retrieval mechanism envisaged for the short term was the GREMAS code, which was automatically generated during the database construction operation.

The 1980s was a decade of activity in this field; by the end of this period two novel public systems had been inaugurated. These are the Markush DARC system of

[⊗] Abstract published in *Advance ACS Abstracts,* August 1, 1996.

S0095-2338(95)00173-9 CCC: $12.00

Derwent Information Ltd., operated by Questel SA and INPI (the French Patent Office), and the MARPAT system of Chemical Abstracts Service.[33−35] They continue in service today, with modifications since their first implementation, and thus reflect continuing development. The Sheffield work undoubtedly influenced these systems; it remains the high-performance ideal which operational systems can regard as the ultimate in accurate and exact representation and retrieval.

2. THE PROBLEMS OF GENERICS

The devices which the authors of chemical patents and their agents use to create variety in their claims and to ensure that the coverage of the patents is as wide as possible have already been described—Dethlefsen *et al.*[24,25] characterized them as the following class-constituting mechanisms:

- Substituent variation—relating to the possible substituents at a fixed position of a partial structure, e.g., "phenyl substituted in the para position by F, Cl, or Br"
- Position variation—relating to multiple positions of attachments of substituents, e.g., "monochlorophenyl"
- Frequency variation—relating to the possible frequencies of partial structures, e.g., "phenyl substituted in the para position by −(CH2)n-Cl; $n = 1-3$"
- Homology variation—relating to the possible combinations of structural features within the limits determined by the intension of expressions like "cycloalkyl" or "phenyl substituted in the para position by Cl-3 alkyl"

The specification of instances of homology variation, i.e., of Homologous Series Identifiers, is often accompanied by numerical or other indications of qualifying restrictions, as illustrated by the following examples:

- number of atoms within a chain
- branching of a chain
- branching and point of attachment of a chain
- number, kind, and position of multiple bonds
- ring size
- number of rings
- pattern of ring fusion

Examples of these are expressions such as

"Cl-8 *n*-alkyl"
"primary dodecanol"
"*n*-dodecadienyl"
"unsaturated unbranched C8-hydrocarbyl chain"
"six-membered heterocycle containing one nitrogen, oxygen, or sulfur atom"

Further features may be superimposed on these, for instance, nesting, in which one variable may be defined in terms of other as yet undefined variables, as in where R_1 is NR_7R_8.

The complexities of the database structures are one side of the problem; the obverse aspect concerns the queries which may be posed against a database of such structures. The queries may be specific structures, the inclusion of which within one or more generic database structures being the criterion to be determined ("inclusion" is defined more closely below); they may be partial structures, i.e., substructures, with or without generic components, inclusion within one or more of the database structures being the criterion, or they may be generic structures, for which more complex matching criteria with other generic structures, again discussed below, may be required.

Table 1. The Scientific Achievements of the Project

A.	The Representation—GENSAL
B.	The ECTR (the Internal Representation) and the Interpreter
C.	Matching Criteria for Generics
D.	Derived Fragment Screens and Search Operations
E.	The Reduced Chemical Graph
F.	The Bubble-Up
G.	The Refined Search—the "Atom-by-atom" Level Search

Central to the issue of searching and matching is the fact that what may appear in specific terms in a query may be satisfied by a generically described moiety in the database structure or *vice versa*. Thus a 2-chloro substituted pyrimidine group may satisfy an element of a query which requires a halo-substituted aromatic heterocycle. Clearly, establishing such equivalences poses problems of some magnitude, and a substantial body of theory is needed to ensure a firm scientific basis in order to ensure database and query description facilities and search algorithms which operate consistently and accurately. Dethlefsen *et al.*[24,25] developed the underlying conceptual framework for the relationship between what appears in patents and what can be represented accurately in the formal structure languages which are necessary in today's systems. In this framework, the syntax, semantics, and pragmatics of chemical structure languages are elaborated, together with the meaning relations between various representation languages which support the development of data structures and algorithms and allow transparent and consistent retrieval of generic structures in response to user queries. The framework also provides guidance toward efficient implementation of search procedures.

In this paper, the principal aspects of the design and implementation of a system reflecting the design requirements are reviewed, together with the theoretical underpinning for the design. Table 1 identifies the principal elements of scientific achievement in respect to the many important issues of representing generic structures, translating them faithfully into a variety of searchable representations, and many other considerations outlined below.

A. GENSAL—the Representation Language. The design of GENSAL, the language devised to describe generic structures for database creation and search purposes, takes full account of the mechanisms which enable the structures to be defined. It also reflects the need to deal with the numeric qualifiers often associated with homologous series identifiers as parameters, e.g., the range of permitted carbon atoms in an alkyl substituent, as in "Cl-8 n-alkyl". The original design of GENSAL envisaged 13 parameters. This number was readily extended in the implementation undertaken at IDC. The GENSAL language has undergone little modification since its initial design; the fact that it is based on a context-free grammar, similar to the framework which underlies modern programming languages, means that such extensions as are necessary for a particular applications context can be readily and systematically incorporated.

We now point out the correspondence between generic chemical structure representations and the combinatorial libraries which are the focus of much attention today. With the exception of homology-variation, in essence, the inclusion of generic nomenclatural terms such as "heterocyclic ring

system", "multiply unsaturated alkyl chain, branched or unbranched", etc., combinatorial libraries are reflections of generic structures exhibiting the more straightforward forms of variation, i.e., substituent-variation, position-variation, and frequency-variation. Immediately, much of the theoretical framework developed for the inherently much more difficult instance of generic structures becomes relevant to the issue of management of the information content of combinatorial libraries; the data structures and manipulations reflecting search algorithms for generics also become relevant.

In another respect, we ourselves did not consider the issue of registration of generic structures seriously, other than to conclude that if the need to determine identity between two database structures were to arise, it could be determined by search. We also specifically ruled out any question of registration of the individual substances covered by the generic structure, for obvious reasons in the case of generic structures involving homology variation. If there were a need to register combinatorial libraries, then, as Dethlefsen has pointed out,[40] this may be done by registering the tabular form of the library. The individual specific substances may then also be identified for registration purposes by devising descriptions in which the cardinal numbers of the individual variables identify the individual substances involved, subject always to the need to identify the recurrence of individual members of the libraries by search, since substances may not always be represented in the same fashion. Thus, in one instance, a substituent may be given as PhR_1, where $R_1 = -CH_2Cl$, $-CH_2OMe$, on the one hand, or again as $PhCH_2R_1$, where $R_1 = Cl$, OMe. Search against the database would be required to identify such recurrences of individual entities.

B. The ECTR (the Internal Representation) and the Interpreter. The Extended Connection Table Representation (ECTR) is the internal representation, the internal counterpart to GENSAL, the external language. This is a tree-structured representation, from which all subsequent data structures are derived. Searchable representations such as fragment screens and ring screens as well as reduced graphs are derived from this ECTR. Unlike GENSAL, the ECTR did see change and extension, as greater insights into the needs and opportunities for translation between representations were gained, or as the disadvantages of earlier designs were seen more clearly.

The ECTR comprises structural information, positional and multiplicity information, and logical information. The structural information to be represented comprises the diagram of the invariant part of the molecule, diagrams defined as instances of variables, and line notations included in textual definitions. The positional and multiplicity information may be derived from graphical sources, e.g., the line attaching a variable group to the center of a ring or ring system, and the possible values of multipliers. Logical information includes indications of what components are in AND or in OR relations to one another. Furthermore, some patents exclude certain combinations of variables in cases where these have already been the subject of earlier publications, and include statements such as "if R_1 is chloro or bromo, then R_2 is not hydrogen". These too are provided for in GENSAL.

The interpreter is an essential component of the software, enabling translations to be made properly and accurately from the external representation to the ECTR and thence to other descriptions. It is thus analogous to the compiler of a high-level programming language. While the GENSAL representation is well suited to human preparation and scrutiny, the several internal representations, including the ECTR, serve various aspects of manipulation and search, including the generation of several levels of screens. These screens include both fragment and ring screens as well as the reduced chemical graph representation. Even in 1980, the need for a variety of screens was envisaged, preferably reflecting different aspects of structural features, and hence to some extent orthogonal to one another.

In addition, automatic text analysis procedures to extract descriptions of generic structures from English-language patent abstracts[37,38] and from the full texts of patents[39] have been developed and show high performance levels.

C. Matching Criteria for Generic Queries. The process of matching queries, which may be substructures, single specific structures, generic substructures, and full generic structures, against substantial databases of generic structures is manifestly expensive in computing resources, both in terms of the initial capital investment in screen record generation as well as in the search process. The hierarchy of screen types exhibits a spectrum of cost and performance levels such that in the beginning small screen sets enable nonrelevant generic structures in the database to be eliminated rapidly, while toward the end of the process smaller numbers of database structures are searched in a much more computationally intensive manner in the refined search, using much more data per structure on average.

It is important to be clear about the criteria for matching queries against database structures, given the variety of possible query types noted above. These are as follows:[25]

- Identity—an exact match exists between query and database structure
- Strict inclusion—every member of the query structure is included in the database structure, but at least one database structure is not included in the query
- Subsumption—includes the two above conditions
- Intersection—at least one structure is common to query and database structures
- Community—includes the conditions of identity, strict inclusion, and intersection
- Strangeness—none of the above conditions holds

Given that the search representations other than the ECTR are degenerate, in the sense that they describe isolated fragments or features only of the query and database structures, special provisions need to be taken to fulfill the search requirements accurately and completely. These are detailed below, for instance, the two-part vector for fragment screens and the need for determining exhaustively all isomorphisms between query and structure, so that all possible mappings at the reduced graph level can go forward to the refined search level.[31]

D. Derived Fragment Screens and Search Operations. The screens chosen for the search operations are similar to those widely used in the screening phase of substructure searches of specific structures databases, as first described by Adamson *et al.*,[41] and extended by Graf et al. within the BASIC system,[42] and by Dittmar et al. in the CAS ONLINE system.[43] They comprise Augmented Atom, Atom Sequence, and Bond Sequence fragment types, a subset of those used in CAS ONLINE, together with special ring fragments. The

augmented atoms describe a central atom, the atoms to which it is connected and the bonds which connect them, at varying levels of description as appropriate to the frequencies of the species in the database and other considerations. Atom and bond sequence fragments describe linear sequences of lengths four, five, and six atoms, subject again to frequency criteria.

The fragments need to reflect the environments from which they are generated, particularly as regards whether they originate from a fixed (i.e., invariant) component of the generic structure or from a variable part. They are differentiated, as they are being generated, according to whether they are derived from within a partial structure (i.e., an Intra-PS fragment) or span several partial structures (i.e., Inter-PS fragments). Search routines use these fragments in the form of two-part records; one part indicates the presence of a fragment in an invariant feature of the structure (MUST screens), while the optional fragments (MAY screens) are used in search as the logical union of the MUST fragments with the MAY fragments (POSS).[16]

A facility identified during studies of ring-screening capabilities has proved to be generally useful in these operations. This is the Bubble-Up (see below), and can be used to manipulate the records of Intra-PS and Inter-PS fragments and other characterizations correctly by means of a logical matrix, and in a manner which is independent of the partitioning which the authors of a patent may have chosen to indicate.

An essential feature of screen generation is that screens are generated not only from those parts of the structures and queries which are described in terms of atoms and bonds, but also from generic components, i.e., from those which are Homologous Series Identifiers. The methods developed for screen generation permit the use either of numerical parameter values, if these are given in the patents, or of default values which characterize the generic radicals, these default values deriving from the underlying theory. The screens generated from generically described components are those which contribute to the description of all of the potential specific instances of the component.

Highly complex problems are posed by use of seemingly innocent phrases such as "R_n and R_m may combine to form a ring .....".

Downs *et al.*[20−23] examined the requirements for ring screening of generic structures; in the event, the existing criteria for the identification of rings within generics which excluded Homologous Series Identifiers were inadequate for the task, and this aspect of theory was successfully extended to take account of the much more complex requirements and resulted in the definition of the Extended Set of Smallest Rings (the ESSR). The ring screens categorize the structure in terms of its rings and their size, composition, and fusion characteristics. In other respects they are treated in the same way as the atom and sequence-based fragments mentioned earlier.

E. The Reduced Chemical Graph. It has already been noted that further and more powerful screen searching than that provided by fragment screens alone would be necessary in order to achieve practicable search performance. A powerful representation, which can be adapted to operate at a number of levels of differentiation, is provided by the Reduced Chemical Graph.[19] The idea of such graphs is well-established in chemical information practices; it is invoked in the creation of chemical names, where it is reflected in the analysis of complex substances into ring and nonring components and similarly in linear notations. The idea involves the fragmentation of structures into ring systems and nonring systems and the description of these both individually and in their relations to one another. Thus nomenclature conveniently divides a structure into these components (apart from details such as carbonyl groups, etc., which may be incorporated into the description of the ring system), and provides a means for naming the ring systems and acyclic components separately and of indicating the relations among them. In this instance, the notion is much the same. Figure 1 illustrates the relation between a generic substance and its reduced chemical graph in which cyclic and acyclic components are differentiated. Note that the inclusion of a hydrogen atom as one of the variables for the R group means that this node become optional; this is indicated by means of a dotted line in the reduced graph. Figure 2 illustrates a simple generic structure.

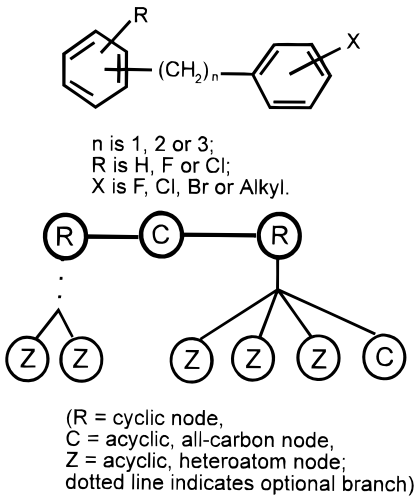

Figure 1. A generic structure and its reduced graph representation.

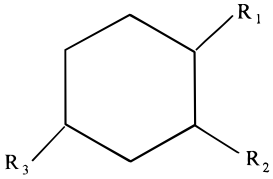

Figure 2. A simple generic structure. R_1 and R_2 are 1-6C alkyl or alkenyl or can combine to form a methylenedioxy or ethylenedioxy ring. R_3 is either methyl or ethyl, substituted by chlorine.

Furthermore, the extent of differentiation in the nodes can be varied at will; acyclic components can be described as such or, alternatively, as consisting exclusively of carbon atoms or of heteroatoms, so that the variety of types and of derived structures is increased. Most usefully in the present context is the fact that a representation which bridges the difference between components described in terms of specific and generic radicals can be devised, in that specific-derived parameters are associated with those nodes which derive from specific components, and the parameters of the generic components, either as given in the patent or as the default values, are assigned to nodes describing homologous series identifiers. The value of these descriptions will be evident in discussion of the Refined Search below.

The reduced chemical graphs are derived by means of complex algorithms which must take account, among other factors, of whether components are optional or otherwise.

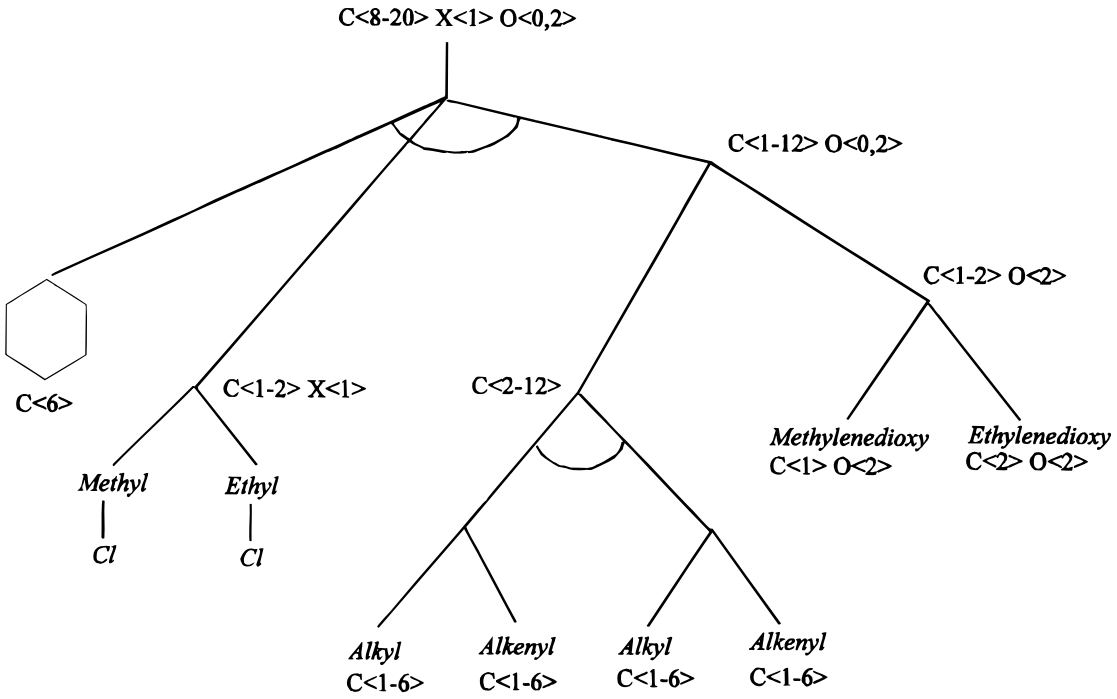

Figure 3. Bubble-up of molecular formulas.

Nodes may be optional either if such an indication is given in the description, or if one of the list of variables is a hydrogen atom. These graphs are examples of AND/OR trees, logical structures which are widely applied in computation. They can be used in the bubble-up process, which was developed by Downs *et al*. while examining means of expressing logical relationships among the ring fragments mentioned above.

F. The Bubble-Up. The bubble-up process operates on the AND/OR tree, and enables features of the structures to be aggregated so that searchable representations, including those which take the form of reduced chemical graphs, can be conveniently produced at a wide range of different levels of detail. The bubble-up process thus provides multiple levels of search, each involving successively more information, analogous to a graded sequence of filters with successively finer meshes. The process was devised during consideration of ring features of generics, where, as noted above, various complex combinations of ring features may result both from rings or ring systems included as variables or as a result of options in choosing combinations of variables which combine to form other ring systems. In fact, the bubble-up is a much more general process, and it can be used to aggregate any structural features associated with nodes of reduced graphs. Figure 3 illustrates the process for the molecular formulas of invariant and alternative nodes in the simple example structure. The process ensures that the aggregation of features accurately reflects the logical relationships of the structure, either as fixed values for those features of the structure which are invariant or as ranges of values for those which are alternative or optional. Thus the invariant part of the structure shown here involves a six-carbon ring with a chloromethyl substituent, accounting for the values C_7Cl_1 in the aggregate. A further carbon atom, as a minimal value, is accounted for by the choice between the required combination of alkyl or alkenyl groups and the methylenedioxy or ethylenedioxy groups. The methylene group, the minimal contributor to the carbon atom count, involves a single C atom, and thus results in an overall minimum of C_8 in the formula. Again, the dioxy ring is optional, hence the alternative for the value of oxygen atoms is zero or two. The determining factor is the AND and OR designation of the arcs of the graph; an AND arc (shown here as an arc across the branch) implies addition of atom counts, while an OR arc implies ORing of the values, resulting in a single value (as for the chlorine atom), in ranges (as for the carbon atoms), or in separate discrete values (as for the oxygen values).

The process of aggregation can be brought up to the root of the tree, as shown here; in other circumstances the process may be halted at intermediate positions in the tree, e.g., at those levels which correspond to nodes of the reduced graph and thus provide for further levels of flexibility in the efficiency of matching routines.

The process is general and can also be applied to the accumulation of fragment and ring screens, provided only that the intracomponent screens are strictly associated with the reduced graph nodes, while intercomponent screen features are handled at the appropriate higher level.

G. The Refined Search—the "Atom-by-Atom" Level Search. This level of search is invoked after all possible screen search operations have been applied. It involves atom-to-atom, atom-parameter, and parameter-parameter searching, at the ultimate level of detail possible. The Ullman algorithm,[44] is used here. It has been extensively evaluated and developed by Willett *et al.*[45−47] for the detection of subgraph isomorphism with a wide variety of types of graphs. The Ullman algorithm has been shown to be highly efficient, and special requirements need to be met here, not least the need to identify all isomorphisms or partial isomorphisms between query and database structure graphs. This requirement arises from the constantly narrowing search process, in which reduced graphs are used as the stage after the fragment screen search, and reduced graphs at successively finer levels of detail are applied successively. At the penultimate and earlier stages of the refined search, all of

the possible correspondences beteen nodes of the reduced graphs of query and database structures need to be maintained, since some valid eliminations or inclusions may be achieved only at the final stage of detail, as described below.

The refined search involves combinations of generic and specific-derived parameters; it is at this level that the systemic nature of these parameters as descriptors of features of structural diversity come most clearly into focus. Consider, for instance, the stage in the process at which two reduced graph components for aliphatic nodes are being compared, one specific, shown below

$$-CH{=}C(CH_3)-CH{=}CH_2$$

and one generic, 4−8C alkadienyl, and described by the following parameters

$$C\langle 4-8\rangle E\langle 2\rangle T\langle 0-\rangle Z\langle 0\rangle$$

i.e., 4−8 carbon atoms (C⟨4−8⟩), two ethylenic bonds (E⟨2⟩), optional ternary branch points (T⟨0−⟩), and no heteroatoms (Z⟨0⟩).

The specific-derived parameters for the group shown in Figure 4 above are

$$C\langle 5\rangle E\langle 2\rangle T\langle 1\rangle Z\langle 0\rangle$$

The parameter values for the specific fragment and for the generically described component are clearly compatible with one another, and thus a match can be determined to exist at this, the final level.

Validation of the correctness of operation of the search routines has been carried out, using a test database compiled in the early years of the project, which has also confirmed the powerful performance of the array of search methods. With the exception of some aspects of frequency variation, the performance has proved to be gratifyingly high.

CONCLUSION

We sought, in the work described here, to establish a rational foundation for considering the requirements for storing and searching generic chemical structures and establishing the representations, data structures, and algorithms, including the all-important search routines, necessary to support the information needs of users of chemical information of this kind.

In large measure we have accomplished these goals and have extended the specific understanding of the area substantially. The work has been validated, in part by internal evaluation in part by external implementation, at least in regard to the database creation aspect, through the development of database creation and translation routines undertaken at IDC. It has undoubtedly influenced the development and continued refinement of today's operational services by providing methods for procedures which the other systems achieve but perhaps less elegantly.

The implementation of the GENSAL-based database creation system by International Documentation for Chemistry (IDC) shortly before its untimely closure represented a major achievement and a *tour de force* on the part of Dr. Guenter Stiegler (now of BASF) and his associates. The purpose at that time was to provide search facilities via the well-understood and widely used GREMAS code. This required extensions to the GENSAL language, in particular, the definition of GENSAL parameters, in order to achieve the degrees of structural differentiation of which the GREMAS code has been capable since its inception, and subsequent automatic translation from the input representation into the appropriate GREMAS screens.

In summary, the area, as expected, proved to be fertile ground for innovative research, throwing up new and valuable ideas, some with potential for application outside of the immediate area for generic chemical structures.

ACKNOWLEDGMENT

We gratefuly acknowledge the assistance provided by the following scientific contributors to the project: J. M. Barnard, L. Carruthers, W. Dethlefsen, G. M. Downs, Val Gillet, E. V. Krishnamurthy, A. Ling, A. von Scholley, G.-U. Schwartz, P. Venkataram, S. M. Welford, and J. V. Wood. The principal funding bodies for the project were British Library R&D Department, Chemical Abstracts Service, Derwent Publications Ltd., Commission of the European Community, International Documentation in Chemistry m.b.H., the U.K. Department for Education, and the University of Sheffield. We also thank the principal external contributors, viz., Dr. Winfried Dethlefsen, Mr. Monty Hyams, Dr. Arthur Kolb, Dr. Ernst Meyer, Dr. Guenter Stiegler, and Dr. Claus Suhr. We also thank a referee for pointing out a problem relating to the original method of dealing with individual entities in combinatorial libraries.

REFERENCES AND NOTES

(1) Rowland, J. F. B. *Information Transfer and Use in Chemistry*; Final Report of the Chemical Information Review Committee, London, British Library R&D Dept., Report No. 5385, 1978.
(2) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Generic Structure Storage and Retrieval. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 264−270.
(3) Meyer, E. *Topological Search for Classes of Compounds in Large Files—even of Markush Formulas—at Reasonable Machine Cost*, in *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E. Eds.; New York, Wiley: 1974; pp 105−122.
(4) Meyer, E.; Schilling, P.; Sens, E. *Experiences with Input, Translation and Search in Files containing Markush Formula Representations, Computer Handling of Generic Chemical Structures*; Barnard, J., Ed.; Aldershot, Gower, 1984; pp 83−95.
(5) Harsdorf, E. von; Dethlefsen, W.; Suhr, C. *Derwent's CPI and IDC's GREMAS: Remarks on their Relative Retrieval Power with Regard to Markush Structures*. *Computer Handling of Generic Chemical Structures*; Barnard, J., Ed.; Aldershot, Gower, 1984; pp 99−105.
(6) Fugmann, R.; Braun, W.; Vaupel, W. GREMAS—a new Method of Classification and Documentation in Organic Chemistry. *Nachr. Dok*. **1963**, *14*, 179−190.
(7) Rössler, S.; Kolb, A. The GREMAS System—an Integral Part of the IDC System for Chemical Documentation. *J. Chem. Doc*. **1980**, *10*, 128−134.
(8) Koller, H. R.; Marden, E.; Pfeffer, H. *The HAYSTAQ System, Past, Present and Future*. In *Proceedings of International Conference on Scientific Information*; Washington, D.C., 1958, NAS/NRC, 1959; Vol. 2, pp 1143−1179.
(9) Krishnamurthy, E. V.; Lynch, M. F. Analysis and Coding of Generic Chemical Structures in Chemical Patents. *J. Inf. Sci*. **1981**, *3*, 75−79.
(10) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 1. Introduction and General Strategy, *J. Chem. Inf. Comput. Sci*. **1981**, *21*, 148−150.
(11) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 2. GENSAL: A Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci*. **1981**, *21*, 151−161.
(12) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 3. Chemical Grammars and their Role in the Manipulation of Chemical Structures. *J. Chem. Inf. Comput. Sci*. **1981**, *21*, 161−168.
(13) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. Part 4. An

Extended Connection Table Represenation (ECTR) for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160−164.
(14) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Towards Simplified Access to Chemical Structure Information in the Patent Literature. *J. Inf. Sci.* **1983**, *6*, 3−10.
(15) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57−66.
(16) Welford, S. M.; Ash, S.; Barnard, J. M.; Carruthers, L.; Lynch, M. F.; Scholley, A. von *The Sheffield University Generic Chemical Structures Reserach Project*, In *Computer Handling of Generic Chemical Structures*; Barnard, J., Ed.; Aldershot, Gower, 1984; pp 130−158.
(17) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 6. An Interpreter Program for the Generic Structure Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66−70.
(18) Gillet, V. J.; Welford, S. M.; Lynch, M. F.; Willett, P.; Barnard, J. M.; Manson, G.; Thomson, J. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118−126.
(19) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 8. Reduced Chemical Graphs and their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126−137.
(20) Downs, G. M.; GIllet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 9. An Algorithm to find the Extended Set of Smallest Rings in Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207−214.
(21) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 10. Assignment and Logical Bubble-up of Ring Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215−224.
(22) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Theoretical Aspects of Ring Perception and Development of the Extended Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187−206.
(23) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172−187.
(24) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 233−253.
(25) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 12. Principles of Search Operations involving Parameter Lists: Matching-relations, User-defined Match Levels, and Transition from the Reduced Graph to the Refined Search. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 253−260.
(26) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260−270.
(27) Holliday, J. D.; Gillet, V. J.; Downs, G. M.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 14. Algorithmic Generation of Fragment Descriptors for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 453−462.
(28) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 15. Generation of Topological Fragment Descriptors from Nontopological Representation of Generic Structure Components. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 369−377.
(29) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. *Searching a Full Generics Database*. In *Chemical Structures, 2*; Warr, W., Ed.; Springer Verlag: Berlin, 1993; pp 87−103.
(30) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F.; Dethlefsen, W. Evaluation of the Screening Stages of the Sheffield Research Project on Computer Storage and Retrieval of Generic Chemical Structures in Patents. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 38−46.
(31) Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 16. The Refined Search: An Algorithm for Matching Components of Generic Chemical Structures at the Atom-bond Level. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1−7.
(32) Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents, Part 17. Evaluation of the Refined Search. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 659−662.
(33) Shenton, K.; Norton, P.; Fearns, E. A. *Generic Searching of Patent Information*, in *Chemical Structures—the International Language of Chemistry*; Warr, W., Ed.; Springer: Berlin, 1988; pp 169−178.
(34) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Storage and Retrieval Capability, Part 1, Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145−155.
(35) Ebe, T.; Sanderson, K. A.; Wilson, P. S. The Chemical Abstracts Service Generic Chemical (Markush) Storage and Retrieval Capability. Part 2. The MARPAT File. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31−36.
(36) Stiegler, G.; Maier, B.; Lenz, H. *Automatic Translation of GENSAL Representations of Markush Structures into GREMAS fragment codes at IDC*. In *Chemical Structures 2*; Warr, W., Ed.; Springer Verlag: Berlin, 1993; pp 105−114.
(37) Chowdhury, G. G.; Lynch, M. F. Automatic Interpretation of the Texts of Chemical Patent Abstract, Part 1, Lexical Analysis and Categorization. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 463−467.
(38) Chowdhury, G. G.; Lynch, M. F. Automatic Interpretation of the Texts of Chemical Patent Abstract, Part 2, Processing and Results. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 468−473.
(39) Kemp, N. Ph.D. thesis, Sheffield University, 1995; Manuscript in preparation.
(40) Dethlefsen, W. Personal communication.
(41) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153−157.
(42) Graf, W.; Kaindl, H. K.; Warszawski, R. The Third BASIC Fragment Search Dictionary. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 177−181.
(43) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Miller, J. A.; Koch, B. The CAS ONLINE Search System. I. General System Design and Selection, Generation and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *33*, 93−102.
(44) Ullman, J. R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.
(45) Willett, P.; Wilson, T.; Reddaway, S. F. Atom-by-Atom Searching Using Massive Parallelism. Implementation of the Ullman Subgraph Isomorphism Algorithm on the Distributed Array Processor. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 225−233.
(46) Brint, A. T.; Mitchell, E.; Willett, P. *Substructure Searching in Files of Three-Dimensional Chemical Structures*. In *Chemical Structures. The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 131−144.
(47) Downs, G. M.; Lynch, M. F.; Willett, P.; Manson, G. A.; Wilson, G. A. Transputer Implementation of Chemical Structure Search Algorithms. *Tetrahedron Comput. Methodol.* **1988**, *1*, 207−217.

CI950173L